Master's Thesis

Trend surface models estimation with outliers

Andrea Meilán Vila

Master in Statistical Techniques

Academic year 2015-2016

II

Master's Thesis Proposal

Título en galego: Estimación de modelos de superficie de tendencia con datos atípicos
Título en español: Estimación de modelos de superficie de tendencia con datos atípicos
English title: Trend surface models estimation with outliers
Modalidad: Modalidad A
Autora: Andrea Meilán Vila, Universidad de Santiago de Compostela
Directora: Rosa M. Crujeiras Casais, Universidad de Santiago de Compostela
Breve resumen del trabajo: En los modelos de superficie de tendencia (modelos de regresión con dependencia espacial donde las variables explicativas son las coordenadas geográficas donde se recogen las observaciones) se suele utilizar la estimación mediante mínimos cuadrados iterados para aproximar los coeficientes del modelo. Sin embargo, este procedimiento es muy poco robusto ante la presencia de datos atípicos en la respuesta. En este trabajo, se plantea el método de estimación, donde se introduce una variable artificial (pseudo-respuesta, versión suavizada de la respuesta original) en el proceso de estimación, con el objetivo de mitigar el efecto de los datos atípicos, en el contexto de datos espacialmente dependientes. También se realiza un estudio de simulación y una aplicación a datos reales.

Doña Rosa M. Crujeiras Casais, profesora contratada doctora del área de Estadística e Investigación Operativa de la Universidad de Santiago de Compostela informa que el Trabajo Fin de Máster titulado

Trend surface models estimation with outliers

fue realizado bajo su dirección por doña Andrea Meilán Vila para el Máster en Técnicas Estadísticas. Estimando que el trabajo está terminado, da su conformidad para su presentación y defensa ante un tribunal.

En Santiago de Compostela, a 1 de Julio de 2016.

La directora:	La autora:
Doña Rosa M. Crujeiras Casais	Doña Andrea Meilán Vila

Contents

Abstract

Resumen

Los modelos de superficie de tendencia tratan de relacionar las observaciones de un proceso espacial que varía de forma continua (proceso geoestadístico) con las coordenadas geográficas en las que dichas observaciones son tomadas, usando modelos de regresión. La diferencia fundamental entre estos modelos y los clásicos de regresión lineal, es que los errores de regresión presentan una estructura de dependencia, que suele ser desconocida, pero debe ser incluida en el modelo.

La estimación de los modelos de superficie de tendencia solo fue abordada en el caso de procesos estadísticos "sencillos", bajo la suposición de normalidad. Sin embargo, incluso bajo esa premisa distribucional, las muestras observadas o los propios procesos pueden presentar ciertas complejidades, como la presencia de datos atípicos, que pueden distorsionar los resultados obtenidos por procedimientos inferenciales.

En este trabajo, se revisan las técnicas exploratorias existentes en la literaura para la detección de datos atípicos y posteriormente, se propone un procedimiento de estimación de modelos de superficie de tendencia que mitigue el efecto de estas observaciones anómalas, introduciendo una variable artificial (pseudo-datos, una versión suavizada de la respuesta original). Se realiza un estudio de simulación y una ilustración con datos reales.

Abstract

Trend surface models try to relate the observations of a spatial process which varies continuously (geostatistic process) with the geographic locations in which those observations are taken, using regression models. The fundamental difference of these models with respect to classical linear regression is that the regression errors presents a dependence structure, which is usually unknown, but which must be included in the model.

Trend surface estimation has only been discussed in the case of "simple" statistical processes, under the assumption of normality. However, even under this distributional premise, the observed samples or own process itself may present certain complexities, as the presence of outliers, which may distort the results obtained by inferential procedures.

In this work, a review of exploratory tools for detecting spatial outliers existing in the literature is initially performed. A procedure for estimating trend surface models which mitigates the effect of these anomalous observations is proposed, introducing an artificial variable (pseudo-data, a smoothed version of the response). Some simulations and an illustration with real data are also presented.

Introduction

This work is framed in the context of spatial stochastic processes, which consists of collections of random variables indexed on a certain domain of $\mathbb{R}^d$, with a well-defined joint distribution. In this setting, the nature of the variable index (which can vary continuously, discretely or randomly), provides a common classification of spatial processes (see Cressie (1993)), being geostastistic a collection of tools and methods developed under the premise of continuous variation of the index. This type of processes (spatial processes indexed continuously, or geostatistics processes) are the data generating mechanism in many applied sciences such as geology, hydrology or environmental sciences.

In all these fields, the observed data tend to exhibit an important feature: close observations tend to be more similar than observations which are far apart. Therefore, such observations cannot be treated as independent and the dependence structure should be taken into account in any descriptive or inferential procedure. In particular, from the perspective of regression models (trend surfaces), the dependence structure should be considered and properly introduced into the model.

Trend surface models (see Diggle and Ribeiro (2007) for a exhaustive description), try to relate the observations of a spatial process which varies continuously (geostatistic process) with the geographic locations in which those observations are taken, using regression models (usually linear). In these models, two sources of variability can be distinguished (see Cressie (1993)): a regression function or trend which would gather large-scale variability and the error term, which would represent small-scale variability. One of the fundamental differences of these models with the classical linear regression is that the errors present a dependence structure, which is usually assumed to be intrinsically stationary or second-order (when working with Gaussian processes, both stationarity conditions are equivalent). That dependence structure is usually unknown but should be included in the model, through the covariogram (if the process is second-order stationary) or the variogram (if the process is intrinsically stationary). See Cressie (1993), Cressie and Hawkins (1980) and Journel and Huijbregts (1978), among others. Estimation of the variogram is preferred to estimation of the covariogram, because it avoids a previous estimation of the trend. In addition, it should be noted that the inclusion of dependence structures is not only relevant for estimating trend surfaces, but also for prediction via spatial interpolators such as kriging.

The problem of trend surface estimation can be solved through least squares tools (where pilot estimations of the variogram are used) or maximum likelihood (the estimation of the parameters of the trend and of the dependence are approached jointly, under the assumption of normality), as it is described in Diggle and Ribeiro (2007). Even so, the trend surface estimation has only been discussed in the literature in the case of "simple" statistical processes, under the assumption of normality. However, even under this distributional premise, observed samples may present certain complexities. In this work we will approach the trend surface estimation when the observations of the process have characteristics that make their treatment more complex, specifically the presence of outliers.

The study of spatial data outliers was approached by exploratory tools, as Moran Scatterplot (see Anselin (1996)) and through observation of the variogram cloud (see Cressie (1993)). The aim of these tools is to identify the possible outliers to remove them before applying any statistical procedure. However, the performance of these tools is far from satisfactory in practice, and requires the intervention of the practitioner. The effect of the presence of outliers or how to handle them in trend surface estimation has not been considered in the statistical literature. After a thorough analysis of the descriptive

tools, we will try to analyze the effect of the presence of outliers in surface trend models and propose an appropriate estimation procedure.

This work is organized as follow. In Chapter 1, a brief description of the univariate geostatistics is realized: different distributional properties of the spatial stochastic processes are revised as well as the estimation of the dependence structure (variogram and covariogram, as required).

Chapter 2 presents the concept of spatial outlier and different coefficients to detect global or local patterns of spatial association, as the Moran I and the Geary's c, and their local counterparts, among others. The review of exploratory tools for detecting spatial outliers existing in the literature, as Moran Scatterplot and variogram cloud, is also performed. A simulation study and an illustration with real data of these techniques are carried out.

Given that existing exploratory tools do not allow a correct identification of spatial outliers, in Chapter 3 we propose a new procedure for estimating trend surface models which mitigates the effect of these anomalous observations. The idea would be to combine the trend surface models estimation using iterative least squares (taking into account the dependence structure) with the use of pseudo-data coming from a previous smoothing of the observed sample (see Akritas (1996) or Cristobal et al. (1987)). Some simulations are performed to ckeck the correct performance of the proposed procedure. The method is also illustrated with real data.

Chapter 1

A brief background on geostatistics

In this chapter, the main concepts and results about geostatistics will be introduced, constituting the basis of this document. Different types of stationarity conditions, which are usually in practice assumed, are revised as well as the estimation of the dependence structure: variogram and covariogram, as required.

As it was mentioned in the Introduction, spatially dependent data and specifically, geostatistical data appear in a variety of applied fields such as geology, hydrology or environmental sciences. Its applications have expanded original mining applications to include modeling soil properties, ground water studies, rainfall precipitation, public health, among others. The generating mechanism of such data is given by a spatial stochastic process:

$$\{Z(\mathbf{s})|\mathbf{s} \in D\}, \tag{1.1}$$

consisting of collections of random variables indexed in a particular domain $D \subset \mathbb{R}^d$ (observation region) with a well defined joint distribution. The nature of the index provides a classification of spatial processes: geostatistical, lattice data and point patterns (see Cressie (1993)). Geostatistics names a collection of tools and methods developed under the premise of continuous variation of the spatial index.

For each location $\mathbf{s}$, $Z(\mathbf{s})$ is a unidimensional or multidimensional random variable. Consider n locations $\{\mathbf{s}_1, \ldots, \mathbf{s}_n\}$ on the region D. The set of random variables corresponding with those locations will be represented by $\{Z(\mathbf{s}_1), \ldots, Z(\mathbf{s}_n)\}$ and $\{z(\mathbf{s}_1), \ldots, z(\mathbf{s}_n)\}$ will denote a realization of this set.

To emphasize the randomness, (1.1) is sometimes written as $\{Z(\mathbf{s}, \omega)|\mathbf{s} \in D, \omega \in \Omega\}$, where $(\Omega, \mathcal{F}, \mathbb{P})$ is a probability space (Ω is the sample space which is not empty, the $\sigma-$algebra $\mathcal{F}$ is a family of subsets of Ω, containing Ω and the empty set, which is closed under the formation of complementary and joints and countable intersections. The probability $\mathbb{P}$ is an application that assigns to each element of $\mathcal{F}$ a number in $[0, 1]$ such that $\mathbb{P}(\emptyset) = 0$ and $\mathbb{P}(\Omega) = 1$, fulfilling that, if $\{A_i, i = 1, \ldots, n\} \subset \mathcal{F}$, with $A_i \cap A_j = \emptyset$, $i \neq j$, then $\mathbb{P}(\cup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} \mathbb{P}(A_i)$). Consequently, the realization $\{z(\mathbf{s})|\mathbf{s} \in D\}$ would correspond to a particular value of ω. Matheron (1962) calls the quantity $z(\cdot)$ a regionalized variable in order to stand out the continuous spatial nature of the index set D.

Usually, a random process has a deterministic and a random (erratic) part, that is to say, $Z(\mathbf{s}) = m(\mathbf{s}) + \varepsilon(\mathbf{s})$ where $m(\cdot)$ usually denotes the deterministic (not random) part and represents large-scale changes, while $\varepsilon(\cdot)$ denotes the random component and shows the local behavior or small-scale evolution. The error process $\varepsilon(\cdot)$ is assumed to be zero-mean, so that the whole trend is captured by $m(\cdot)$. Therefore, $m(\cdot)$ would give the first order structure (mean) and $\varepsilon(\cdot)$ explains the second-order part (as the mean is zero, what contributes is the part of covariance). The large-scale spatial structure of the process (global, over the entire region) is represented by the mean function while the small-scale structure (local, highly localized region) is explained by the covariogram or the variogram, as will be shown below.

1.1. Distributional properties

The stochastic behaviour of (1.1) can be explained through the finite dimensional distributions

$$F_{\mathbf{s}_1,\ldots,\mathbf{s}_n}(z_1,\ldots,z_n) = \mathbb{P}(Z(\mathbf{s}_1) \leq z_1,\ldots,Z(\mathbf{s}_n) \leq z_n)), \quad n \geq 1, \quad \mathbf{s}_i \in D, \;\; i = 1,\ldots,n.$$

which must satisfy Kolmogorov's conditions of symmetry (F remains invariant when z_j and s_j are subjected to the same permutation) and consistency ($F_{\mathbf{s}_1,\ldots,\mathbf{s}_{k+l}}(z_1,\ldots,z_k,\infty,\ldots,\infty) = F_{\mathbf{s}_1,\ldots,\mathbf{s}_k}(z_1,\ldots,z_k)$).

In any case, we will focus our attention in Gaussian spatial processes, that is to say, those processes whose finite-dimensional distribution is normal or Gaussian. These processes are important for two reasons: firstly, under the assumption of normality, prediction, estimation and distribution theory are easier; moreover, many small order effects (possibly non-Gaussian) are asymptotically Gaussian (central limit theorem), as it can be seen in Lindgren (1976).

When considering Gaussian processes, and bearing in mind the decomposition of the process in large and small variability sources, then there is no need to control the whole distribution and just the first two moments are required. In fact, in most practical applications, available information does not allow to infer higher order moments. The first moment is the expectation, which is defined as

$$\mathbb{E}[Z(\mathbf{s})] = m(\mathbf{s}), \quad \mathbf{s} \in D.$$

If this exists for all $\mathbf{s} \in D$, it is called the trend (sometimes drift) of the random process. The three second order moments considered in geostatistics are the variance or second order moment of $Z(\mathbf{s})$ with respect to $m(\mathbf{s})$,

$$\sigma^2(\mathbf{s}) = \mathrm{Var}[Z(\mathbf{s})] = \mathbb{E}\{[Z(\mathbf{s}) - m(\mathbf{s})]^2\},$$

the covariance of two random variables $Z(\mathbf{s})$ and $Z(\mathbf{s}')$ (which is just function of $\mathbf{s}$ and $\mathbf{s}'$),

$$C(\mathbf{s},\mathbf{s}') = \mathrm{Cov}[Z(\mathbf{s}), Z(\mathbf{s}')] = \mathbb{E}\{[Z(\mathbf{s}) - m(\mathbf{s})][Z(\mathbf{s}') - m(\mathbf{s}')]\},$$

and the variogram which is defined as the variance of the difference process:

$$2\gamma(\mathbf{s},\mathbf{s}') = \mathrm{Var}[Z(\mathbf{s}) - Z(\mathbf{s}')].$$

To perform an exploratory data analysis, the traditional numerical summaries (mean, median, mode, standard deviation, range, interquartile range) reduces the data to a few numbers. This approach may not be suitable for a geostatistical analysis, provided that location information is ignored. Methods to explore spatial data include the scatterplot of the index variable versus the variable, means and medians of rows and columns, variogram cloud (see Section 2.2.2.) among others.

In general, it is not possible to make statistical inference from a single realization $\{z(\mathbf{s}_1),\ldots,z(\mathbf{s}_n)\}$. To enable statistical inference in this approach, it is essential to introduce additional assumptions about the process $\{Z(\mathbf{s})|\mathbf{s} \in D\}$, assuming stationarity of some kind: strict, second-order (or weak) and instrinsic stationarity, that will be defined below. The type of stationarity which is assumed indicates the type of statistical inference that can be made with the probabilistic model: if the random process $\{Z(\mathbf{s})|\mathbf{s} \in D\}$ is second-order stationary or intrinsically stationary, the dependence structure of the stochastic process will be specified by the covariogram in the first case and by the variogram in the second case.

Definition 1.1. *A spatial process $\{Z(\boldsymbol{s})|\boldsymbol{s} \in D\}$ is strictly stationary if its finite-dimensional distribution is invariant to translations, that is, $F_{\boldsymbol{s}_1,\ldots,\boldsymbol{s}_n}(z_1,\ldots,z_n) = F_{\boldsymbol{s}_1+t,\ldots,\boldsymbol{s}_n+t}(z_1,\ldots,z_n)$, $t \in \mathbb{R}^d$, $(\boldsymbol{s}_i + t) \in D \;\forall i = 1,\ldots,n$, and $n \in \mathbb{N}$.*

This condition is not usually checked in real data examples. For this reason, other less restrictive hypotheses are assumed. Since geostatistics (in Gaussian processes) is based on the first two moments, it is sufficient to assume that these two moments exist and limit the hypothesis of stationarity for the first two moments. The following definitions, among many other results related to spatial data, can be found in Cressie (1993), a classical reference in this topic.

Definition 1.2. *A spatial process is second-order stationary if* $\mathbb{E}[Z(\boldsymbol{s})] = m$, $\forall \boldsymbol{s} \in D$ *and* $C(\boldsymbol{s}, \boldsymbol{s}') = C(\boldsymbol{s} - \boldsymbol{s}')$, $\forall \boldsymbol{s}, \boldsymbol{s}' \in D$.

That is, second-order structure is just a function of the difference vector between the locations where the observations are taken. The function $C(\cdot)$ is called covariogram or covariance function.

Such processes satisfy that $\sigma^2(\mathbf{s}) = \text{Var}[Z(\mathbf{s})] = C(0) = \sigma^2$, $\forall \mathbf{s} \in D$, that is, the variance of a second-order stationary process is finite and independent of the spatial location $\mathbf{s}$. Note that the previous definition implies that the mean of the process is constant. So, in the presence of a trend (that is, $m(\cdot)$ is not constant), this trend is usually estimated first and stationarity is checked over the residual process.

Note that if the last condition in Definition 1.2 is replaced by $\text{Cov}[Z(\mathbf{s}), Z(\mathbf{s}')] = \text{C}(\|\mathbf{s} - \mathbf{s}'\|)$, where $\|\cdot\|$ represents the euclidean norm, then the second-order stationary process is also isotropic. That is, second-order stationarity means that dependence between two observations is only a function of the difference vector between the locations where the observations are taken, while isotropy goes further, considering that dependence is only a function of distance, ignoring direction. Otherwise, the process will be anisotropic.

Second order stationarity does not imply strict stationarity. Conversely, a stationary process may not be second-order stationary, as its first two moments may not be defined. To illustrate this feature, consider a process $Z(\mathbf{s})$ defined in the following way. At each location $\mathbf{s}$, an observation $z(\mathbf{s})$ is drawn at random from the Cauchy distribution. For this particular case, none of the moments (mean and variance) exist because the corresponding integrals do not converge absolutely, but the distribution is exactly the same at all points and is translation invariant. Hence, the process is strictly stationary, but does not satisfy the definition of second-order stationarity given above.

For a Gaussian random process, second-order stationarity implies strict stationarity, because a Gaussian process is characterized by its mean and its covariance function. Of course, strict stationarity implies second-order stationarity whenever F yields a finite second moment.

It is possible that the assumption of second-order stationarity is too strong. As given by Matheron (1971), the form of stationarity implied by the intrinsic hypothesis is essentially second-order stationarity, not for the process $Z(\mathbf{s})$, but for the difference, $Z(\mathbf{s} + \mathbf{u}) - Z(\mathbf{s})$:

$$\mathbb{E}(Z(\mathbf{s} + \mathbf{u}) - Z(\mathbf{s})) = 0, \quad \forall \mathbf{u} \in \mathbb{R}^d, \ \forall \mathbf{s}, \mathbf{s} + \mathbf{u} \in D \tag{1.2}$$

$$\text{Var}(Z(\mathbf{s} + \mathbf{u}) - Z(\mathbf{s})) = 2\gamma(\mathbf{u}), \quad \forall \mathbf{u} \in \mathbb{R}^d, \ \forall \mathbf{s}, \mathbf{s} + \mathbf{u} \in D. \tag{1.3}$$

Definition 1.3. *Suppose* $\{Z(\boldsymbol{s}) | \boldsymbol{s} \in D\}$ *satisfies (1.2) and (1.3). Then* $Z(\cdot)$ *is said to be intrinsically stationary (or, equivalently, to satisfy the intrinsic hypothesis).*

The function $2\gamma(\cdot)$ has been called variogram (and $\gamma(\cdot)$ has been called semivariogram) by Matheron (1962), although earlier appearances can be found in the literature. It has beeen called a structure function by Yaglom (1957) in probability theory and a mean squared difference by Jowett (1952) in time series.

Later, it will be seen that average squared differences are used to estimate $2\gamma(\cdot)$, however, modeling $\gamma(\cdot)$ is all that is needed. That is why $\gamma(\cdot)$ is sometimes called "variogram" by abuse of notation, as in this document.

Note that, considering that dependence is only a function of distance then the intrinsic process is isotropic. Otherwise, the process is anisotropic. This hypothesis does not require that $\text{Var}[Z(\mathbf{s})]$ is finite and independent of the spatial location $\mathbf{s}$, but usually it is.

Second order stationarity implies intrinsic stationarity while the reciprocal is true only when the variance is constant and finite. The best known example of a process that is intrinsic, but not second-order stationary, is the Wiener-Levy process or, as it is often called, Brownian motion (see Mörters and Peres (2010)).

1.2. Variogram and covariogram

Suppose that the random process $\{Z(\mathbf{s})|\mathbf{s} \in D\}$ is second-order stationary or intrinsically stationary, the dependence structure of the stochastic process will be specified by the covariogram in the first case and by the variogram in the second case.

It can be seen that $\gamma(\cdot)$ is a symmetric function since $\gamma(\mathbf{u}) = \gamma(-\mathbf{u}), \forall \mathbf{u} \in D$. Furthermore, $\gamma(\mathbf{u}) \geq 0$ and $\lim_{\|\mathbf{u}\|\to\infty} \gamma(\mathbf{u})/\|\mathbf{u}\|^2 = 0$, $\forall \mathbf{u} \in D$, that is to say, the variogram should increase more slowly than $\|\mathbf{u}\|^2$. Moreover, a variogram is valid if it satisfies the conditionally negative definite property:

$$\sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j \gamma(\mathbf{s}_i - \mathbf{s}_j) \leq 0, \forall \mathbf{s}_1, \ldots, \mathbf{s}_n \in D \text{ and } \forall a_1, \ldots, a_n \in \mathbb{R} \text{ such that } \sum_{i=1}^{n} a_i = 0.$$

It is always true that $\gamma(0) = 0$, but if $\gamma(\mathbf{u}) \to c_0 \neq 0$, as $\mathbf{u} \to 0$, then c_0 has been called the *nugget effect* by Matheron (1962). This is because it is believed that microscale variation (small nuggets) is causing a discontinuity at the origin (this cannot happen for L_2-continuous processes, since these ones satisfy that $\mathbb{E}(Z(\mathbf{s}+\mathbf{u}) - Z(\mathbf{s}))^2 \to 0$, as $\|\mathbf{u}\| \to 0$, $\forall \mathbf{s}, \mathbf{s}+\mathbf{u} \in D$). Hence, if continuity of the phenomenon is expected at the microscale, the only possible reason for $c_0 > 0$ is measurement error (let us just call c_{ME} at the error variance). Given that only observations $\{Z(\mathbf{s}_i), i = 1, \ldots, n\}$ are available and nothing can be said about the variogram at lag distances smaller than $\min\{\|\mathbf{s}_i - \mathbf{s}_j\|\}$, it is not known whether the microscale variation is continuous or not, but Matheron typically makes the assumption that it is not. To model the process at very small scales, Matheron adds a white-noise process (zero mean, constant variance and zero covariance) to a process with continuous sample paths (call the variance of this white-noise process c_{MS}, which represents the nugget effect of the microscale process). Thus, $c_0 = c_{MS} + c_{ME}$. Therefore, c_0 can be included as a parameter in the variogram model, but it is hard to determine its value from data whose separations are too large to give accurate microscale information. Typically, it is determined by extrapolating variogram estimates from lags closes to zero.

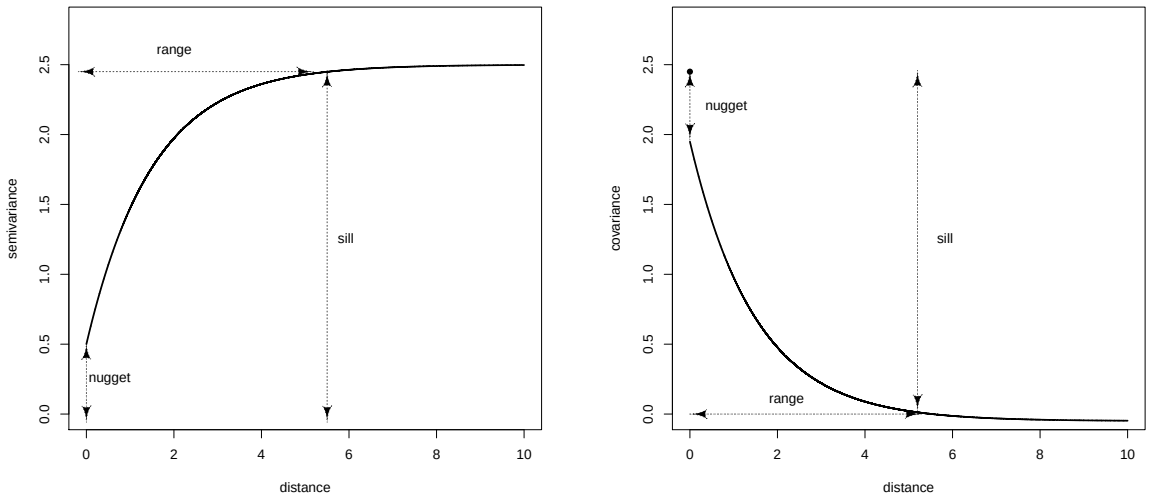

Figure 1.1: A generic variogram (left) and its corresponding covariogram (right) showing the sill and range parameters, with the nugget effect.

Now if $\gamma(\cdot)$ is bounded and there is $\lim_{\|\mathbf{u}\|\to\infty} \gamma(\mathbf{u})$, this limit is called the *sill* and does not always exist. If the process is second-order stationary then the sill coincides with the variance σ^2. When the

semivariogram has nugget effect, the difference $\sigma^2 - c_0$ is called *partial sill*. If σ^2 is the sill, the *range* (if it exists) is a real value r such that if $\|\mathbf{u}\| \geq r$, then $\gamma(\mathbf{u}) = \sigma^2$ (it is equivalent to say that the variables $Z(\mathbf{s})$ and $Z(\mathbf{s}+\mathbf{u})$ are uncorrelated). When the process is second-order stationary then the *asymptotic range* can be defined as a real value r' such that if $\|\mathbf{u}\| \geq r'$, then $\gamma(\mathbf{u}) \geq c_0 + 0.95(\sigma^2 - c_0)$ (see Chilès and Delfiner (1999)). Figure 1.1 (left) shows a generic variogram identifying the role of the parameters. There are several parametric models of the variogram which will be introduced later. These are valid models which satisfy the assumptions mentioned.

The covariogram or covariance function characterizes the dependence structure for second-order stationary stochastic process $\{Z(\mathbf{s})|\mathbf{s} \in D\}$, and has the following properties. $C(\cdot)$ is a symmetric function, and it holds that $C(\mathbf{0}) = \mathrm{Var}(Z(\mathbf{s})) \geq 0$, $\forall \mathbf{s} \in D$. In addition, by Cauchy-Schwarz inequality, it can be seen that $|C(\mathbf{u})| \leq C(\mathbf{0})$. It must be positive definite,

$$\sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j C(\mathbf{s}_i - \mathbf{s}_j) \geq 0, \ \forall \mathbf{s}_1, \ldots, \mathbf{s}_n \in D \text{ and } \forall a_1, \ldots, a_n \in \mathbb{R}.$$

In this case, the *sill* is the covariance at zero distance, while the *range* is the distance at which covariance reaches zero (could be infinity). The *partial sill* is the limit of the covariance at $\|\mathbf{u}\| \to 0$, from the right. The *nugget* is the sill minus the partial sill. The Figure 1.1 (right) shows a generic covariogram identifying the role of the parameters.

The study of the variogram is more common than the covariance, since intrinsic processes are more general than second-order stationary processes. Furthermore, the estimation of the variogram does not requires knowledge of the process mean or its estimate. If the process is second-order stationary, an estimator of the covariogram can be obtained from an estimator of the variogram just considering that

$$\begin{aligned} \gamma(\mathbf{u}) &= \frac{1}{2}\mathrm{Var}[Z(\mathbf{s}) - Z(\mathbf{s}+\mathbf{u})] \\ &= \frac{1}{2}\mathrm{Var}[Z(\mathbf{s})] + \frac{1}{2}\mathrm{Var}[Z(\mathbf{s}+\mathbf{u})] - \mathrm{Cov}[Z(\mathbf{s}), Z(\mathbf{s}+\mathbf{u})] \\ &= \sigma^2 - C(\mathbf{u}). \end{aligned} \tag{1.4}$$

In addition to the variogram and covariogram, a related function is the *correlogram*. If $C(\mathbf{0}) > 0$, the correlogram is defined as

$$\rho(\mathbf{u}) = \frac{C(\mathbf{u})}{C(\mathbf{0})}.$$

It can be seen that $\rho(\cdot)$ is a symmetric function since $\rho(\mathbf{u}) = \rho(-\mathbf{u})$, $\forall \mathbf{u} \in D$, and that $\rho(\mathbf{0}) = 1$. Furthermore, as a result, from (1.4) it can be seen that

$$\rho(\mathbf{u}) = 1 - \frac{\gamma(\mathbf{u})}{C(\mathbf{0})}.$$

1.3. Estimation of the variogram

In order to provide an estimator of the variogram, a simple option is to consider its empirical counterpart: the empirical variogram. Consider an intrinsic stationary process. Then $\mathbb{E}(Z(\mathbf{s}+\mathbf{u}) - Z(\mathbf{s})) = 0$ and $\mathrm{Var}(Z(\mathbf{s}+\mathbf{u}) - Z(\mathbf{s})) = 2\gamma(\mathbf{u})$. So the problem of estimating the variogram is reduced to estimate $\mathbb{E}[(Z(\mathbf{s}+\mathbf{u}) - Z(\mathbf{s}))^2]$ from a random sample $\{Z(\mathbf{s}_i), i = 1, \ldots, n\}$.

The classical empirical estimator of the variogram proposed by Matheron (1962) is based on the method of moments and is given by

$$2\hat{\gamma}(\mathbf{u}) = \frac{1}{|N(\mathbf{u})|} \sum_{(i,j)\in N(\mathbf{u})} (Z(\mathbf{s}_i) - Z(\mathbf{s}_j))^2, \tag{1.5}$$

where $|N(\mathbf{u})|$ is the number of pairs in $N(\mathbf{u}) \equiv \{(i,j) : \mathbf{s}_i - \mathbf{s}_j = \mathbf{u}\}$. If there are few data pairs whose difference is $\mathbf{u}$, $N(\mathbf{u})$ may be substituted by a region of tolerance $A(\mathbf{u}) \equiv \{(i,j) : \mathbf{s}_i - \mathbf{s}_j \approx \mathbf{u}\}$, where differences $(\mathbf{s}_i - \mathbf{s}_j)$ are close to $\mathbf{u}$ at according to a specified tolerance. The isotropic version of (1.5) only depends on the distance between points and can be represented in two dimensions.

The empirical estimator is unbiased. However, it possesses very poor resistance properties, where the word "resistant" refers to a statistic that is arithmetically stable under gross contamination of the data values. It is badly affected by outliers due the $(\cdot)^2$ term in the addend of (1.5). The variogram cloud (see Cressie (1993)) shows all pairwise distance vectors combined with the squared differences of the observations $Z(\mathbf{s}_1), \ldots, Z(\mathbf{s}_n)$. Cressie (1993), claims that it is difficult to distinguish atypical observations from skewness using the variogram cloud. This is because for a Gaussian process $Z(\cdot)$, $(Z(\mathbf{s}) - Z(\mathbf{s}+\mathbf{u}))^2$ is distributed as $2\gamma(\mathbf{u})\chi_1^2$ since $\mathbb{E}(Z(\mathbf{s}+\mathbf{u}) - Z(\mathbf{s})) = 0$ and $\mathrm{Var}(Z(\mathbf{s}+\mathbf{u}) - Z(\mathbf{s})) = 2\gamma(\mathbf{u})$. Thus, $2\gamma(\mathbf{u})$ is the first moment of a chi-squared random variable on one degree of freedom, which is highly skewed.

If the data are normally distributed, the estimator obtained by the method of moments coincides with that obtained by maximum likelihood. This would indicate that $2\hat{\gamma}(\cdot)$ provides a good approximation of the function $2\gamma(\cdot)$. However, if the data do not show a close behavior to the normal distribution, it may be advisable to use a robust estimator of the variogram.

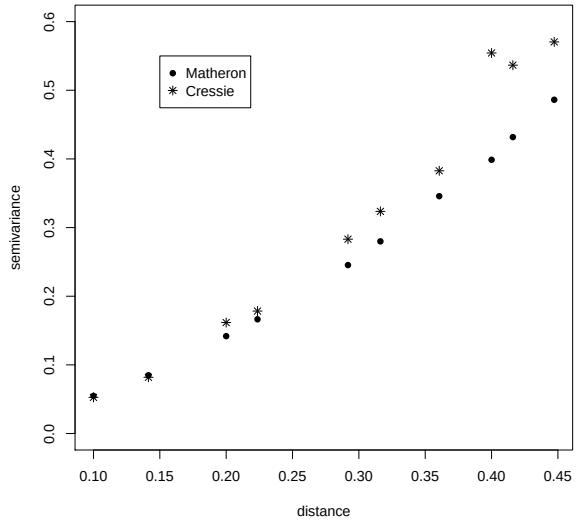

Figure 1.2: Empirical and robust estimators of the variogram. Sample size: 121. Regularly simulated data from a Gaussian spatial process on unit square with exponential covariogram (1.13) with parameters $c_0 = 0$, $c_e = 1$ and $a_e = 2$.

Cressie and Hawkins (1980) present a more robust approach to the estimation of the variogram by transforming the problem to location estimation for an approximately symmetric distribution (to avoid skewness), in particular, the estimation of a center of symmetry. The class of power transformations $\{Z(\mathbf{s}+\mathbf{u}) - Z(\mathbf{s})\}^\lambda$ was chosen. A theoretical study showed that the fourth-root of χ_1^2 has a skewness of 0.08 and a kurtosis of 2.48, compared with 0 and 3 for the Gaussian distribution (see Bowman and Crujeiras (2013) to compare the density function of the fourth-root transformation of a χ_1^2 random variable with the density function of a normal random variable with the same mean and standard deviation). Estimators of location parameters, such as the mean and the median, can then be applied to the $|N(\mathbf{u})|$ transformed differences $\{|Z(\mathbf{s}_i) - Z(\mathbf{s}_j)|^{1/2} : (i,j) \in N(\mathbf{u})\}$. Finally, these estimates are raised to the fourth power, to bring them back to the correct scale, and adjusted for bias. This results

in variogram estimators,

$$2\bar{\gamma}(\mathbf{u}) = \frac{1}{(0.457 + 0.494/|N(\mathbf{u})|)} \left\{ \frac{1}{|N(\mathbf{u})|} \sum_{(i,j)\in N(\mathbf{u})} |Z(\mathbf{s}_i) - Z(\mathbf{s}_j)|^{1/2} \right\}^4 \tag{1.6}$$

and

$$2\tilde{\gamma}(\mathbf{u}) = \frac{1}{B(\mathbf{u})} \left[\text{med} \left\{ |Z(\mathbf{s}_i) - Z(\mathbf{s}_j)|^{1/2} : (\mathbf{s}_i, \mathbf{s}_j) \in N(\mathbf{u}) \right\} \right]^4, \tag{1.7}$$

where med$\{\cdot\}$ denotes the median function and $B(\mathbf{u})$ corrects for bias (asymptotically, $B(\mathbf{u}) = 0.457$).

A simulation is performed to illustrate the bahavior of empirical and robust estimators of the variogram. A sample of size 121 is simulated regularly from a Gaussian spatial process considering the unit square as support and taking the exponential model given in (1.13) as covariance function with $c_0 = 0$, $c_e = 1$ and $a_e = 2$. Note that the range value, 2, is greater than any of the distances in the unit square, leading a strong dependence (see Figure 1.2).

Anisotropy

Before fitting a variogram it should be determined if it can support the hypothesis of isotropy or, if not, what kind of anisotropy exists.

To determine (in an exploratory way) if the isotropy hypothesis can be assumed, an alternative is to build the empirical variograms in several directions and establish if they follow the same pattern. For the latter, one can construct the variogram, under isotropy conditions, and it can be compared with those obtained in different directions. If variograms are different then there will be anisotropy. Empirical variograms (of the same simulated data used in the Figure 1.2) in the directions of 0°, 45°, 90° and 135° are in the Figure 1.3. It can be seen that empirical variograms are different, especially when the directions of 45° and 135° are considered, which indicates the presence of a possible anisotropy.

It is also useful to draw variogram lines, consisting in computing the values of the variogram for different distances in different directions and join distances for which the same values of the variogram would be achieved. If this representation gives rise to concentric circles, the condition of isotropy is supported. If these lines represent concentric ellipses, then there will be geometric anisotropy. In the two-dimensional case, to transform the previous ellipse in a circle in a circle, it would be enough to make the following change of coordinates:

$$\begin{pmatrix} x' \\ y' \end{pmatrix} = \begin{pmatrix} \lambda\cos\theta & \lambda\sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix},$$

where λ is the anisotropy ratio (ratio between the minimum and maximum range) and θ is the direction of maximum range or angle anisotropy.

Some formal tests for isotropy have been proposed in the literature, based on asymptotic results for the empirical variogram (1.5). Lu and Zimmerman (2005) use the spectral density, which requires regularly spaced data. Guan et al. (2004) use sub-sampling to construct an estimate of the covariance matrix and this requires selection of the size of the sub-samples as well as the number of lags. These authors also recommend identifying particular directions of interest. Bowman and Crujeiras (2013) proposed a test of isotropy too. Consider

$$\hat{\gamma}^\star(\mathbf{u}) = \frac{1}{|N(\mathbf{u})|} \sum_{(i,j)\in N(\mathbf{u})} |Z(\mathbf{s}_i) - Z(\mathbf{s}_j)|^{1/2},$$

which is an estimate of $\gamma^\star(\mathbf{u}) = 0.977741\{\gamma(\mathbf{u})\}^{1/4}$ (see Bowman and Crujeiras (2013)). Variograms which are unaffected by a change in angle θ will also lead to absence of effect on the fourth-root scale.

The mean and covariance descriptions of $\hat{\gamma}^{\star}(\cdot)$ provide a quantitative assessment of the evidence for anisotropy. In this case, $\hat{\gamma}^{\star}(\mathbf{u},\theta)$ can be estimated by constructing bin means d_b over a two-dimensional grid of values of $(\mathbf{u},\theta)$. Evidence for anisotropy then rests on comparing an estimate based on $(\mathbf{u},\theta)$ with an estimate based on $\mathbf{u}$ alone. Denoting the vector of bin means by $\mathbf{d}$, fitted values based on smoothing can be expressed as $\hat{\gamma}_1^{\star} = M_1\mathbf{d}$ and $\hat{\gamma}_0^{\star} = M_0\mathbf{d}$, where M_1 and M_0 are smoothing matrices which incorporate distance and angle, and only distance, respectively. Since angle lies on a cyclical scale, this feature should be incorporated into the angle component of the smoothing matrix M_1. This is done with a two-dimensional p-spline basis. Under the assumption of isotropy, the difference between the smooth estimators, $(M_1 - M_0)\mathbf{u}$, has mean 0. Therefore, global evidence for anisotropy can be quantified through the test statistic

$$\mathbf{d}^T M^T \hat{V_0}^{-1} M\mathbf{d} \tag{1.8}$$

where $M = M_1 - M_0$ and V_0 is the covariance matrix for the isotropic case. The use of smoothing matrices means that statistic (1.8) has a χ^2 distribution.

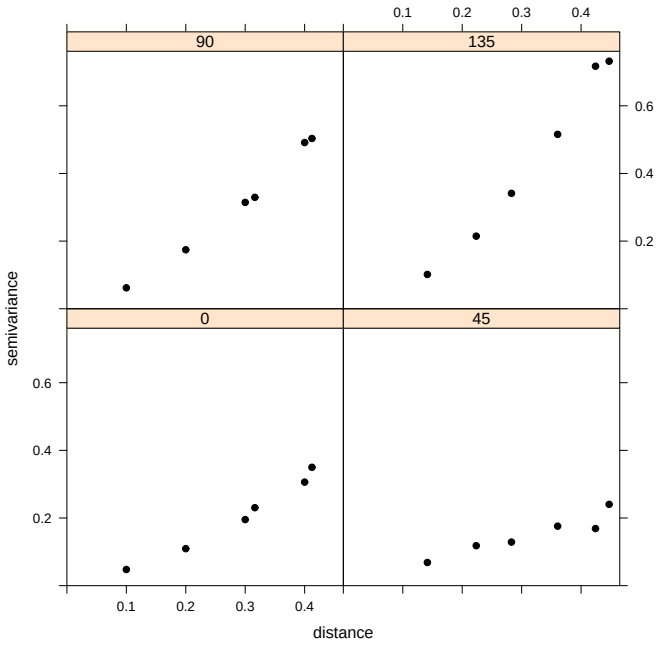

Figure 1.3: Empirical variograms in the directions of 0°, 45°, 90° and 135°. Sample size: 121. Regularly simulated data from a Gaussian spatial process on unit square with exponential covariogram (1.13) with parameters $c_0 = 0$, $c_e = 1$ and $a_e = 2$.

The anisotropy may be of two kinds: geometric and zonal. The geometric anisotropy occurs when the range varies with the direction, while the zonal occurs when there is an additional source of variability in one direction, and therefore the sill depends on the direction. For geometric anisotropy, the estimation of the variogram is reduced to find a matrix A such that $\gamma(\mathbf{u}) = \gamma_0(\|A\mathbf{u}\|)$, $\forall \mathbf{u} \in D$, where $\gamma_0(\cdot)$ is isotropic. Note that, before studying the anisotropy it is necessary to consider that the directional behaviour may result from the process is not stationary, or not even intrinsically stationary.

1.3.1. Isotropic variogram models

The methods described in the previous section are useful for estimating the values of the variogram at certain distances or lags, however, some practical application may require all values of the variogram

function or semivariogram. In addition, the variograms and covariograms must satisfy the properties mentioned above. In particular, semivariogram must be conditionally negative definite.

The empirical variogram is a function that is not defined for all lags and it is not guaranteed that all the properties required for a variogram to be valid are satisfied. In practice, it is calculated and adjusted to a semivariogram model.

Given that $\mathbf{u}$ is a vector and $\gamma(\cdot)$ is a scalar function, $\gamma(\cdot)$ may depend on the distance $u = \|\mathbf{u}\|$ as well as on the orientation. A great simplification is obtained by assuming that dependence structures are functions only of the distance, that is to say, assume isotropy ($u = \|\mathbf{u}\|$). The three basic isotropic models given in Journel and Huijbregts (1978) are the linear, spherical and exponential models. In all these models the parameter c_0 represents the nugget effect.

- The linear model is valid in $\mathbb{R}^d$, $d \geq 1$, and it is defined as

$$\gamma(u;\boldsymbol{\theta}) = \begin{cases} 0, & u = 0, \\ c_0 + b_l u, & u \neq 0, \end{cases} \tag{1.9}$$

  where $\boldsymbol{\theta} = (c_0, b_l)'$, with $c_0 \geq 0$ and $b_l \geq 0$. If the parameter $b_l = 0$, then we are in the presence of an indicative model of a phenomenon without spatial autocorrelation, and is called pure nugget effect model. This model is linear and differentiable on its parameters (other models are nonlinear in its parameters although differentiable on them). The linear variogram has no sill, and so the variance of the process is infinite. Figure 1.4 (full line) is a linear variogram with $c_0 = 0.5$ and $b_l = 3$.

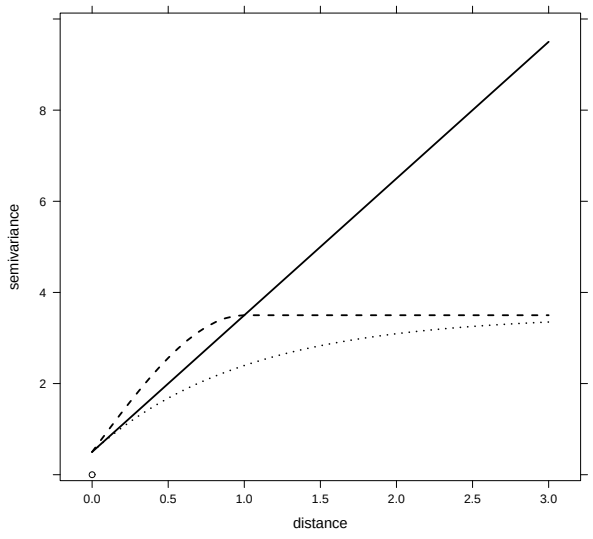

Figure 1.4: Linear (full line), spherical (dashed line) and exponential (dotted line) semivariograms models with $\boldsymbol{\theta} = (c_0, b_l)' = (0.5, 3)$, $\boldsymbol{\theta} = (c_0, c_s, a_s)' = (0.5, 3, 1)$ and $\boldsymbol{\theta} = (c_0, c_e, a_e)' = (0.5, 3, 1)$, respectively.

- The spherical model is valid in $\mathbb{R}$, $\mathbb{R}^2$ and $\mathbb{R}^3$ (but for higher dimensions it fails the non-positive

definiteness condition), and it is defined as

$$\gamma(u;\boldsymbol{\theta}) = \begin{cases} 0, & u = 0, \\ c_0 + c_s\{(3/2)(u/a_s) - (1/2)(u/a_s)^3\}, & 0 < u \leq a_s, \\ c_0 + c_s, & u \geq a_s, \end{cases} \tag{1.10}$$

where $\boldsymbol{\theta} = (c_0, c_s, a_s)'$, with $c_0 \geq 0$, $c_s \geq 0$ (partial sill) and $a_s \geq 0$ (range). It increases from 0 when u is small, levelling off at the constant $c_0 + c_s$ at $u = a_s$. The slope at the origin is equal to $1.5(c_o + c_s)/a_s$. To see the effect of these parameters, it is useful to consider the spherical variogram shown in Figure 1.4 (dashed line) with $c_0 = 0.5$, $c_s = 3$ and $a_s = 1$.

The spherical covariogram corresponding to expression (1.10) is immediately obtainable from (1.4), and is given by:

$$C(u;\boldsymbol{\theta}) = \begin{cases} c_0 + c_s, & u = 0, \\ c_s\{1 - (3/2)(u/a_s) - (1/2)(u/a_s)^3\}, & 0 < u \leq a_s, \\ 0, & u \geq a_s. \end{cases} \tag{1.11}$$

While the spherical model is smooth in the sense of continuous differentiability, it makes the implicit assumption that correlations are exactly zero at all sufficiently large distances. However, in some cases it may be more appropriate to assume that while correlations may become small at large distances, they never vanish completely. The simplest model with this property is the exponential variogram.

- The exponential model is valid in $\mathbb{R}^d$, $d \geq 1$, and it is defined for all $u \geq 0$ by

$$\gamma(u;\boldsymbol{\theta}) = \begin{cases} 0, & u = 0, \\ c_0 + c_e\{1 - \exp(-u/a_e)\}, & u \neq 0, \end{cases} \tag{1.12}$$

where $\boldsymbol{\theta} = (c_0, c_e, a_e)'$, with $c_0 \geq 0$, $c_e \geq 0$ (partial sill) and $a_e \geq 0$ (range). The exponential model reaches the sill asymptotically,

$$\lim_{u\to\infty} c_0 + c_e\{1 - \exp(-u/a_e)\} = c_o + c_e.$$

Practical range is usually taken as the distance at which $\gamma(u) = 0.95(c_0 + c_e)$. The slope at the origin is equal to $(c_o + c_s)/a_s$, which is less than the slope in a spherical variogram with the same parameter range.

Here it is clear that the sill, c_e, and nugget, c_o, play the same role as in the spherical model. However, the range parameter, a_e, is more difficult to interpret. Figure 1.4 (dotted line) is a exponential variogram with $c_0 = 0.5$, $c_e = 3$ and $a_e = 1$.

The corresponding exponential covariogram is defined for all $u \geq 0$ by

$$C(u;\boldsymbol{\theta}) = \begin{cases} c_0 + c_e, & u = 0, \\ c_e\{\exp(-u/a_e)\}, & u \neq 0. \end{cases} \tag{1.13}$$

Other variogram models are the following: Gaussian, wave and power models.

- The Gaussian model is valid in $\mathbb{R}^d$, $d \geq 1$, and it is defined by

$$\gamma(u;\boldsymbol{\theta}) = \begin{cases} 0, & u = 0, \\ c_0 + c_g\{1 - \exp(-u^2/a_g^2)\}, & u \neq 0, \end{cases} \tag{1.14}$$

  where $\boldsymbol{\theta} = (c_0, c_g, a_g)'$, with $c_0 \geq 0$, $c_g \geq 0$ (partial sill) and $a_g \geq 0$ (range). The Gaussian semivariogram approaches its sill asymptotically too. A practical range is used which is equal to the distance at which the semivariogram is equal to 95 % of the sill. The left hand panel of Figure 1.5 (full line) shows a Gaussian variogram with $c_0 = 0.5$, $c_g = 3$ and $a_g = 1$. A modified function can be used with an extra parameter ω replacing the squared exponent in the Gaussian model given in (1.14).

  The corresponding Gaussian covariogram is defined for all $u \geq 0$ by

$$C(u;\boldsymbol{\theta}) = \begin{cases} c_0 + c_g, & u = 0, \\ c_g\{\exp(-u^2/a_g^2)\}, & u \neq 0. \end{cases} \tag{1.15}$$

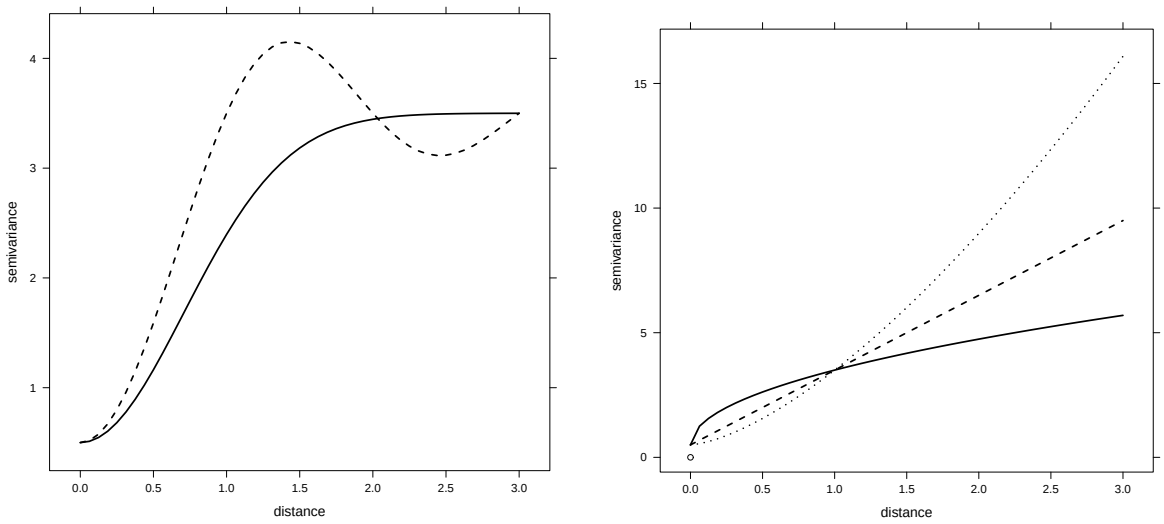

Figure 1.5: The left hand panel shows Gaussian (full line) and wave (dashed line) variograms models with $\boldsymbol{\theta} = (c_0, c_g, a_g)' = (0.5, 3, 1)$ and $\boldsymbol{\theta} = (c_0, c_w, a_w)' = (0.5, 3, 1)$, respectively. The right hand panel shows power semivariograms models with parameters $\boldsymbol{\theta} = (c_0, b_p, \lambda)' = (0.5, 3, 0.5)$ (full line), $\boldsymbol{\theta} = (c_0, b_p, \lambda)' = (0.5, 3, 1)$ (dashed line) and $\boldsymbol{\theta} = (c_0, b_p, \lambda)' = (0.5, 3, 1.5)$ (dotted line).

- A semivariogram model that exhibits negative correlations caused by periodicity of the process is the wave (or hole-effect) model (valid in $\mathbb{R}$, $\mathbb{R}^2$ and $\mathbb{R}^3$). Variations in the depth of the ocean due to the action of the waves on the surface would be a practical example. The semivariogram model is defined as

$$\gamma(u;\boldsymbol{\theta}) = \begin{cases} 0, & u = 0, \\ c_0 + c_w\{1 - a_w \sin(u/a_w)/u\}, & u \neq 0, \end{cases} \tag{1.16}$$

where $\boldsymbol{\theta} = (c_0, c_w, a_w)'$, with $c_0 \geq 0$, $c_w \geq 0$ (partial sill) and $a_w \geq 0$ (range). To interpret these parameters, it can be useful to consider the wave variogram shown in the left hand panel of Figure 1.5 (dashed line) with $c_0 = 0.5$, $c_w = 3$ and $a_w = 1$.

The corresponding wave covariogram is defined by

$$C(u;\boldsymbol{\theta}) = \begin{cases} c_0 + c_w, & u = 0, \\ c_w\{a_w \sin(u/a_w)/u\}, & u \neq 0. \end{cases} \tag{1.17}$$

- The power model is valid in $\mathbb{R}^d$, $d \geq 1$, and it is defined by

$$\gamma(u;\boldsymbol{\theta}) = \begin{cases} 0, & u = 0, \\ c_0 + b_p u^{\lambda}, & u \neq 0, \end{cases} \tag{1.18}$$

where $\boldsymbol{\theta} = (c_0, b_p, \lambda)'$, with $c_0 \geq 0$, $b_p \geq 0$ and $0 \leq \lambda < 2$. The power variogram has no sill, so the variance of the process is infinite. The linear variogram is a special case of the power model. It presents different behaviors at the origin depending on the value of λ. The right hand of Figure 1.5 (full line) is a power variogram with $c_0 = 0.5$, $b_p = 3$ and $\lambda = 0.5$ while the dotted line of the same panel is a power variogram with $c_0 = 1$, $b_p = 3$ and $\lambda = 1.5$. If $\lambda = 1$, then the model is linear (see Figure 1.4 (full line) and the right hand panel of Figure 1.5 (dashed line)).

An example of a random process in $\mathbb{R}$ which has variogram $2\gamma(u) = |u|^{\lambda}$, $0 \leq \lambda < 2$, is the fractional Brownian motion (see Mandelbrot and Van Ness (1968)). To prove it, it would be sufficient taking into account that the fractional Brownian motion is defined by its stochastic representation

$$B_H(t) = \frac{1}{\Gamma(H+1/2)} \left(\int_{-\infty}^{0} [(t-s)^{H-1/2} - (-s)^{H-1/2}] dB(s) + \int_0^t (t-s)^{H-1/2} dB(s) \right),$$

with parameter $H = \lambda/2$, and its moments. This is a zero-mean Gaussian process with $B_H(0) = 0$, stationary increments and $\mathrm{Cov}(B_H(t), B_H(t')) = (1/2)\{|t|^{2H} + |t'|^{2H} - |t-t'|^{2H}\}$, $0 < H < 1$ and $0 < t \leq t'$. Note that $H = 1/2$ corresponds to the standard Brownian motion.

1.3.2. Variogram model fit

It is not guaranteed that the different variogram estimators, $\hat{\gamma}$, $\bar{\gamma}$ and $\tilde{\gamma}$ (empirical, robust and the correction for bias, respectively), satisfy the conditionally negative definite property. As a consequence, it is possible that some spatial predictions obtained from these estimators present negative variances. To avoid this problem, the estimators $\hat{\gamma}$, $\bar{\gamma}$ and $\tilde{\gamma}$ are replaced by a parametric model, which satisfies conditionally negative definite assumption.

The idea is to look for a valid variogram which represents (as accurately as possible) the spatial dependence between the observed data. In general, if

$$\{\gamma : \gamma(\cdot) = \gamma(\cdot;\boldsymbol{\theta}), \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$$

is the set of valid variograms, then an element of the set that best fits the sample variogram must be found. Two approaches are the most widely used: maximum likelihood and least squares. The maximum likelihood requires assumptions about the distribution of $\{Z(\mathbf{s}_i), i = 1, \ldots, n\}$ (usually Gaussian), while criteria based on ordinary, generalized and weighted least squares do not require these assumptions to estimate $\boldsymbol{\theta}$. Note that, maximum likelihood allows an approximation of the trend and of the error simultaneously.

In general, as variograms estimated by likelihood methods are not based on empirical variograms, there are differences between these and the estimators based on least squares.

Maximum likelihood

Maximum likelihood method could be considered in order to obtain a valid parametric variogram estimator. This estimation procedure relies on the Gaussian assumption, whereas the least squares method only depends on the asymptotic second-order structure of the process.

Consider a Gaussian process, for simplicity denoted as $\mathbf{Z} \sim N(X\boldsymbol{\beta}, \Sigma(\theta))$, where $\mathbf{Z}$ refers to the vector of observations (n-dimensional), X is a matrix of covariables (size $n \times q$, with $q < n$), $\boldsymbol{\beta}$ is a q-dimensional vector and $\Sigma(\theta)$ is the covariance matrix of the observations. Given that $\mathbf{Z} \sim N(X\boldsymbol{\beta}, \Sigma(\theta))$ then its density function will be

$$f(z) = (2\pi)^{-n/2}\det(\Sigma(\theta))^{-1/2}\exp\left\{-\frac{1}{2}(\mathbf{Z}-X\boldsymbol{\beta})'\Sigma(\theta)^{-1}(\mathbf{Z}-X\boldsymbol{\beta})\right\},$$

and the negative log-likelihood will be

$$\mathcal{L} = \frac{n}{2}\log(2\pi) + \frac{1}{2}\log(\det(\Sigma(\theta))) + \frac{1}{2}\left\{(\mathbf{Z}-X\boldsymbol{\beta})'\Sigma(\theta)^{-1}(\mathbf{Z}-X\boldsymbol{\beta})\right\}.$$

A maximum likelihood estimator (MLE) is a vector $(\hat{\boldsymbol{\beta}}, \hat{\theta})$ that maximizes $\mathcal{L}$. Generally a MLE must be found by numerical optimization routines, as Newton-Raphson. Thus starting values and convergence criteria must be selected. Besides, the MLE may present a serious bias, although this problem can be mitigated using a restricted maximum likelihood (REML) approach (an example of MLE bias is obtained when $\varepsilon = \sigma^2 I$).

REML is an approach that produces unbiased estimators for these special cases, and produces less biased estimates than ML in general. The REML method consist in finding $n - \text{rank}(X) = n - r$ linearly independent vectors $\mathbf{a}_1, \ldots, \mathbf{a}_{n-r}$ such that $\mathbf{a}_i' X = \mathbf{0}$ for all $i = 1, \ldots, n-r$. It must find the maximum likelihood estimate of $\theta \in \Theta$ using $w_1 = \mathbf{a}_1\mathbf{Z}, \ldots, w_{n-r} = \mathbf{a}_{n-r}\mathbf{Z}$ as data.

In general, REML estimation provides better results than ML estimation, as it gives rise to estimators with smaller deviations (for samples with few data). Moreover, REML estimation is widely used in geostatistics.

Least squares

Suppose that the variogram $\gamma(\cdot)$ is estimated in a finite set of distances, and that the goal is to fit a parametric model $\gamma(\cdot;\boldsymbol{\theta})$ (the vector $\boldsymbol{\theta}$ contains the nugget effect, the sill and the range). Suppose that the method of moments estimator $\hat{\gamma}$ has been used and let be $\hat{\boldsymbol{\gamma}}$ the vector which contains the estimated values. $\boldsymbol{\gamma(\theta)}$ will denote the vector of values obtained by the parametric model in the same values of u.

- Ordinary least squares (OLS), in which $\boldsymbol{\theta}$ is chosen as the vector that minimizes the expression $\{\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma(\theta)}\}'\{\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma(\theta)}\}$ (where $'$ denote the transpose of a matrix).
- Weighted least squares (WLS), in which $\boldsymbol{\theta}$ is chosen as the vector that minimizes the expression $\{\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma(\theta)}\}'W^{-1}\{\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma(\theta)}\}$. W is a diagonal matrix which elements of the diagonal are the variances of $\hat{\boldsymbol{\gamma}}$.
- Generalized least squares (GLS), in which $\boldsymbol{\theta}$ is chosen as the vector that minimizes the expression $\{\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma(\theta)}\}'V(\boldsymbol{\theta})^{-1}\{\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma(\theta)}\}$. $V(\boldsymbol{\theta})$ denotes the covariance matrix of $\hat{\boldsymbol{\gamma}}$, which depend on $\boldsymbol{\theta}$.

Determining $V(\boldsymbol{\theta})$ for minimizing the expression $\{\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma(\theta)}\}'V(\boldsymbol{\theta})^{-1}\{\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma(\theta)}\}$ is not always easy. Cressie (1993) proposed as finding $V(\boldsymbol{\theta})$ in the case of the classical estimator.

Chapter 2

Spatial outliers

The identification of outliers is important since the behavior of these data is unusual with respect to the whole dataset and their inclusion in statistical procedures (or their consideration without further refinements) may compromise the derived conclusions. They are different definitions of outliers. Hawkins (1980) defined an outlier as an observation which deviates so much from other observations as to arouse *suspicions* that it was generated by a different mechanism. An alternative definition of outlier was given by Beckman and Cook (1983), who defined an outlier as a contaminant or a discordant observation, where a *discordant* observation refers to any observation that appears *surprising* or *discrepant* to the investigator, and a *contaminant* is any observation that is not a realization from the target distribution. Barnett and Lewis (1994) established an outlier as an observation (or subset of observations) which appears to be *inconsistent* with the remainder of that set of data. It is a matter of subjective judgement on the part of the observer whether or not he picks out some observation (or set of observations) for scrutiny.

Since the goal of this chapter is the review of techniques for the detection of spatial outliers, specific features of spatial data need to be defined first. Spatial data is a collection of spatially referenced objects which have two categories of dimensions of special interest: spatial and non-spatial. Spatial attributes of a spatially referenced object include location (geographic coordinates, for example), while non-spatial attributes include the observation taken. Furthermore, according to Shekhar et al. (2003), a spatial neighborhood of a spatially referenced object is a subset of the spatial data based on a spatial dimension, generally, location. These neighborhoods may be defined from spatial attributes using spatial relationships, such as distances. More precisely, if $\mathbf{s}_i \in D$ is a location (spatial attibute) and $Z(\mathbf{s}_i)$ is the observation taken in this location (non-spatial attribute), then based on a distance u, its neighborhood may be defined as $\{Z(\mathbf{s}_j) : j \in D,\ \|\mathbf{s}_i - \mathbf{s}_j\| \leq u\}$.

Consequently, a spatial outlier is defined as a spatially referenced object whose non-spatial attribute values are *significantly different* from those of other spatially referenced objects in its spatial neighborhood. More precisely, a spatial outlier is a local instability (in values of non-spatial attributes) or a spatially referenced object whose non-spatial attributes are extreme relative to its neighbours, although they may not be significantly different from all data. Knowledge of the local behaviour as well as detecting spatial outliers is useful in many applications of geographic information systems and spatial databases, including transportation, ecology, public safety, public health, climatology, and location-based services.

The detection of local patterns of spatial association is an important concern in spatial process. In this chapter, local indicators of spatial association (LISA) will be introduced, and how they allow for the decomposition of global indicators, such as Moran's I, into the contribution of each observation. These LISA statistics serve two purposes. On the one hand, they may be interpreted as indicators of clusters, or hot spots. On the other hand, they may be used to assess the influence of individual locations on the magnitude of the global statistic and to identify outliers. An initial evaluation of the distribution of LISA statistics as well as the review of other exploratory techniques to detect outliers, as

the Moran scatterplot (see Anselin (1995)) or the variogram cloud (see Cressie (1993)), are carried out. Some simulations will be done to check these tools. An application to real data will also be performed.

Note that, in this chapter we denote by $\{z_1, \ldots, z_n\}$ the set of observations to simplify the notation.

2.1. Indicators of spatial association

To verify the existence or absence of spatial autocorrelation, different coefficients can be used (the Moran's I is the best known and used in practice). All these coefficients try to test the null hypothesis of no spatial autocorrelation against the alternative hypothesis of spatial autocorrelation. Spatial autocorrelation indicators may have a global or local nature. Global indicators are limited to the hypothesis of spatial autocorrelation in whole territory studied, but do not allow to determine whether the scheme of spatial autocorrelation detected throughout the whole territory is also maintained locally. Local indicators detect the possible presence of spatial autocorrelation in a particular subset. Thus, an index can be obtained for each spatial unit studied, allowing to analyze the degree of individual dependence of each spatial unit relative to the others (positive or negative). In the next sections, both types of indicators will be revised.

2.1.1. Global indicators

Global indicators show the presence or absence of a stable pattern of spatial dependence that is true for the whole dataset. Of all global coefficients, the Moran's I is the most used and is given by the following expression (see Moran (1948); Cliff and Ord (1973)):

$$I = \frac{n}{S_0} \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}\tilde{z}_i\tilde{z}_j}{\sum_{i=1}^{n}\tilde{z}_i^2}, \tag{2.1}$$

where n stands for the number of observations $\tilde{z}_i$[1], w_{ij} is the ij-element of the spatial weights matrix W and S_0 is the sum of all elements in the spatial weights matrix ($S_0 = \sum_i \sum_j w_{ij}$). When the spatial weights matrix is row-standarized such that the elements in each row sum to 1, the expression (2.1) simplifies to

$$I^\star = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}\tilde{z}_i\tilde{z}_j}{\sum_{i=1}^{n}\tilde{z}_i^2}, \tag{2.2}$$

since in this case, the factor S_0 equals n (since each row sums to 1), and the statistic simplifies to a ratio of a spatial cross product to a variance.

Values for Moran's I do not need to be constrained to the interval $[-1, 1]$, however, usually $|I| \leq 1$, unless regions with extreme values of $\tilde{z}_i$ (remember that they are in deviations from their mean) are heavily weighted. In the absence of autocorrelation and regardless of the specified weight matrix, the expectation of Moran's I statistic is $-1/(n-1)$, which tends to zero as the sample size increases. A Moran's I coefficient larger than $-1/(n-1)$ indicates positive spatial autocorrelation, and a Moran's I less than $-1/(n-1)$ indicates negative spatial autocorrelation.

A second popular global index of spatial autocorrelation is Geary's c (see Geary (1954);Cliff and Ord (1981)), which is based on a weighted average of the similarity values observed for all pairs assigning

[1] $\tilde{z}_i$ denotes centred observations.

weights by spatial proximity. This weighted average is scaled by a measure of overall variation around the sample mean. The coefficient is defined as

$$c = \frac{(n-1)}{2S_0} \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(\tilde{z}_i - \tilde{z}_j)^2}{\sum_{i=1}^{n} \tilde{z}_i^2}, \tag{2.3}$$

using the same notation as before. Geary's c ranges from 0 (maximal positive autocorrelation) to 2 (indicating perfect negative spatial autocorrelation). Its expectation is 1 in the absence of autocorrelation and regardless of the specified weight matrix (if the value of Geary's c is less than 1, it indicates positive spatial autocorrelation).

In addition to the previous quantities, there are other coefficients such as Mantel's Γ (see Mantel (1967)), which is defined as

$$\Gamma = \sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij} b_{ij}, \tag{2.4}$$

where a_{ij} and b_{ij} are the elements of two matrices of similarity. Measures of spatial association are obtained by expressing similarities by means of matrices: spatial similarity (for example, the spatial weight matrix) and value similarities. Different measures of value similarity yield different indices for spatial association. For example, using $a_{ij} = \tilde{z}_i \tilde{z}_j$ yields a Moran's measure and setting $a_{ij} = (\tilde{z}_i - \tilde{z}_j)^2$ yields a Geary's index.

Moreover, spatial autocorrelation may be measured as a distanced-based or spatial clustering measure. For the following test, two spatial units are neighbors if they are located at a certain distance d. Getis and Ord (1992) defined the $G(d)$ coefficient as

$$G(d) = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(d)\tilde{z}_i\tilde{z}_j}{\sum_{i=1}^{n}\sum_{j=1}^{n} \tilde{z}_i\tilde{z}_j}, \tag{2.5}$$

where $W(d) = (w_{ij}(d))$ is a spatial weight matrix with ones for all pairs defined as being within distance d of a given i and for all other pairs are zero including the pair of point i to itself. $G(d)$ will not show negative spatial autocorrelation. Note that $G(d)$ does not include the i-th attribute, but there is another version $G^\star(d)$ which also takes into account this attribute as well as neighborhood values.

Moran's I coefficient depends on the difference between each value of the attribute and its mean. However, Geary's $c$oefficient depends on the (absolute) difference between neighboring values of a variable.

Moran's I is a more global measurement and sensitive to extreme values of the attribute, whereas Geary's c is more sensitive to differences in small neighborhoods. However, Moran's I and Geary's c generally result in similar conclusions. Cliff and Ord (1973, 1981) showed that Moran's I is consistently more powerful than Geary's c, so Moran's I is usually prefered.

Moran's I and Geary's c only indicate global clustering. They can not report if these are hot spot (a region where high values cluster together) or cold spots (a region where low values cluster together). However, the $G(d)$ statistic distinguishes between hot spots and cold spots. It identifies spatial concentrations: $G(d)$ is large if high values cluster together, while $G(d)$ is low if low values cluster together.

2.1.2. Local indicators

Local indicators provide a value for each observation, since different patterns may occur in different parts of the region, but an equivalent local coefficient can be calculated for most global measures.

The local form of Moran's I, given in (2.1), for the i-th observation is defined as

$$I_i = \tilde{z}_i \sum_{j=1}^{n} w_{ij}\tilde{z}_j, \quad i = 1, \ldots, n, \tag{2.6}$$

where the observation $\tilde{z}_i$ is in deviation from the mean too, and the summation over j is such that only neighboring values $j \in \{1, \ldots, n\}$ are included. For ease of interpretation, the weights w_{ij} may be in row-standardized form (though this is not necessary), and by convention, $w_{ii} = 0$.

Using the same principles as before, a local Geary's c statistic of (2.3) for each observation may be defined as

$$c_i = \sum_{j=1}^{n} w_{ij}(\tilde{z}_i - \tilde{z}_j)^2, \quad i = 1, \ldots, n, \tag{2.7}$$

using the same notation as before.

Since the Γ index given in (2.4) is a simple sum over i, a local Gamma index for a location i may be defined as

$$\Gamma_i = \sum_{j=1}^{n} a_{ij}b_{ij}, \quad i = 1, \ldots, n. \tag{2.8}$$

Similar to what holds for the global Γ, different measures of value similarity will yield different indices of local association.

The $G(d)$ statistic given in (2.5) for an observation, measures the concentration of the weighted sum of values of the attributes in a subregion of j locations around i in the global region. It is defined as

$$G_i(d) = \frac{\sum_{j=1}^{n} w_{ij}(d)\tilde{z}_j}{\sum_{j=1}^{n} \tilde{z}_j}, \quad i = 1, \ldots, n, \tag{2.9}$$

for $j \neq i$ and where $W(d) = (w_{ij}(d))$ is a spatial weight matrix with ones for all pairs defined as being within distance d of a given i and for all other pairs are zero including the pair of point i to itself. The numerator is the sum of all $\tilde{z}_j$ within d of i but not including $\tilde{z}_i$. The denominator is the sum of all $\tilde{z}_j$ not including $\tilde{z}_i$ (see Getis and Ord (1992)), there are a version of $G_i(d)$ which is denoted by $G_i^\star(d)$ and defined as

$$G_i^\star(d) = \frac{\sum_{j=1}^{n} w_{ij}(d)\tilde{z}_j}{\sum_{j=1}^{n} \tilde{z}_j}, \quad i = 1, \ldots, n, \tag{2.10}$$

for all j. This statistic differs from $G_i(d)$ in that $G_i^\star(d)$ includes the value of the point in its computation. $G_i(d)$ excludes this value and only considers the value of its nearest neighbours against the global average (which also does not inclue the value at i). $G_i^\star(d)$ is the more common of the two statistics because it considers all values.

The first interpretation of LISA is the identification of local spatial clusters. A positive value of I_i indicates a spatial clustering of similar values (either high or low) and a negative value a clustering of dissimilar values (a location with high values surrounded by neighbors with low values, or vice versa). A positive value of $G_i^\star(d)$ indicates a spatial clustering of high values, and a negative value a spatial clustering of low values. The individual Γ_i may be interpreted as indicators of significant local spatial clusters.

The second interpretation of a LISA is a diagnosis for outliers with respect to a measure of global association previously given. The distribution of I_i statistics can provide an indication of outliers, by means of a 2σ rule (percentage of values that lie within a band around the mean in a normal distribution with a width of two standard deviations, 95.45 % of the values). Observations exceeding this threshold are considered. Although this is not a test, it provides useful insight into the special nature of these observations (the identification of outliers may realise in the box plot of the I_i too). Note that this notion of extremeness does not imply that the corresponding I_i are significant in the sense outlined earlier, but only indicates the importance of observation i in determining the global statistic. A diagnostic for outliers can be carried out by comparing the distribution of the Γ_i to Γ/n.

The moments for I_i under the null hypothesis of no spatial correlation can be derived using the principles outlined by Cliff and Ord (1981). A test for significant local spatial correlation may be based on these moments, although the exact distribution of such a statistic is still unknown. The probability distribution may be badly represented by a normal distribution. Alternatively, a conditional randomization approach by permutation (unknown distribution function) may be taken. Given the structure of the statistic in (2.6), it follows that only the quantity $\sum_{j=1}^{n} w_{ij}\tilde{z}_j$ to be computed for each permutation. Following the suggestion by Ord annd Getis (1994), a Bonferroni bounds procedure is used to assess significance. With a α level of 0.05, the individual significance levels for each observation should be taken as $0.05/n$.

The only aspect of equation (2.9) that changes with each permutation is the numerator, since the denominator does not depend on the spatial allocation of observations. This is the same term as the varying part of (2.6). The randomization method applied to (2.6) will yield the same empirical reference distribution as when applied to Getis and Ord $G_i(d)$ and $G_i^{\star}(d)$ statistics. Hence, inference based on this nonparametric approach will be identical for the two statistics. The significance levels (that is, the generated with a permutation approach applied to the I_i statistic will be identical for $G_i(d)$ and $G_i^{\star}(d)$ statistics.

2.2. Exploratory techniques for detecting spatial outliers

Global outliers detection methods ignore the spatial location of each data points while spatial outliers methods separate spatial attributes from non-spatial attributes. According to Anselin (1996), methods of spatial association can be classified in two different groups, depending on the way in which spatial interaction is conceived.

Firstly, based on geographical information, this association could be seen as a covariation between neighbouring observations. Moran scatterplot (see Anselin (1995, 1996)) is a exploratory tool to visualise and identify the degree of spatial instability in spatial association by means of Moran's I. As for Moran's I, the neighborhood structure of a dataset is collected in a spatial weights matrix W, with elements $w_{ij} = 0$ when i and j are not neighbours and non-zero otherwise (as a rule, w_{ii} is assumed to be zero). This tool is based on the interpretation of the Moran's I statistic as a regression coefficient.

Secondly, based on geostatistics, the spatial interaction is conceptualised as a continuous function of a metric distance. The method of choice is the variogram or semivariogram, which is based on the squared difference between values observed at a given distance (see Cressie (1993)).

There are other tools to detect spatial outliers, including the box map, the map of percentiles and the cartogram (see Anselin (2005)).

2.2.1. Moran scatterplot

Firstly, based on geographical information, according to Anselin (1995), the Moran sccaterplot combined with a classical linear regression will be used to detect spatial outliers. Abusing of notation, to explain this exploratory tool, we denote by z_i the standarized observations, $(z_i - \bar{z})/s_z$, $i = 1, \ldots, n$ ($\bar{z}$ and s_z denote the sample mean and standard deviation of the values of the attributes z_i, $i = 1, \ldots, n$, respectively). The Moran scatterplot gives a formal indication of the degree of linear association bet-

ween a vector of normalized observed values z_i, and a weighted average of the neighbouring normalized attributes, $Y_{w,i} = \sum_j w_{ij} z_j$, $i = 1, \ldots, n$, where $w_{ij} > 0$ and $\sum_j w_{ij} = 1$. The k-Nearest Neighbors algorithm will be used to compute the weight matrix w_{ij} and as a consequence, the detected outliers will depend on k. Note that the presence of outliers may also point to problems with the specification of the spatial weights matrix.

Since the z_i are observations in deviations from their mean, and $Y_{w,i}$ is the associated spatial lag, the scatter plot is centred at $(0, 0)$. The four quadrants in the plot represent different types of association between z_i and $Y_{w,i}$. The upper right and lower left quadrants represent positive association in the sense that a observation in a location has similar values to those in its neighborhood. For the upper right quadrant, it is a association between high values (above the mean) while for the lower left quadrant it is between low values (below the mean). The upper left and lower right quadrants correspond to negative association, that is, low values are rounded by high values (upper left) and high values are surrounded by low values (lower right). As mentioned, the pairs $(Y_{w,i}, z_i)$ are given for standardize values, so that outliers may be easily visualized as points further than two units away from the origin.

According to Anselin (1995), the application of regression diagnostics for leverage and residuals to the scatterplot suggest that observations may deserve closer scrutiny. Points in the scatterplot that are extreme with respect to the central tendency reflected by the slope may be outliers in the sense that they do not follow the same process of spatial dependence as the bulk of the other observations. An intuitive indication of outliers can be based on the normalized residuals from the regression of Y_w on z.

The classical linear regression model will be

$$Y_{w,i} = \beta_0 + \beta_1 z_i + \varepsilon_i, \quad i = 1, \ldots, n \tag{2.11}$$

where $\varepsilon_1, \ldots, \varepsilon_n \in N(0, \sigma^2)$ and independent. Consequently, the fitted values $\hat{Y}_{w,i}$ would be

$$\hat{Y}_{w,i} = \hat{\beta}_0 + \hat{\beta}_1 z_i, \quad i = 1, \ldots, n,$$

where $\hat{\beta}_0$ and $\hat{\beta}_1$ are the estimators of β_0 and β_1, respectively. In short, it would have the following prediction errors or residuals,

$$\hat{\varepsilon}_i = Y_{w,i} - \hat{Y}_{w,i} = Y_{w,i} - \hat{\beta}_0 - \hat{\beta}_1 z_i, \quad i = 1, \ldots, n.$$

The idea is to choose the estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ which give the smaller regression residuals. It can be shown, using least squares, that $\hat{\beta}_0 = \bar{Y}_w - \hat{\beta}_1 \bar{z}$, $\hat{\beta}_1 = \sum_{j=1}^n (z_i - \bar{z}) Y_{w,j} / Szz$ and $Szz = \sum_{j=1}^n (z_j - \bar{z})^2$. Therefore,

$$\begin{aligned} \hat{Y}_{w,i} &= \bar{Y}_w + \hat{\beta}_1 (z_j - \bar{z}) = \frac{1}{n} \sum_{j=1}^n Y_{w,j} + \sum_{j=1}^n \frac{(z_j - \bar{z})}{Szz} Y_{w,j} (z_i - \bar{z}) \\ &= \sum_{j=1}^n \left[\frac{1}{n} + \frac{(z_i - \bar{z})(z_j - \bar{z})}{Szz} \right] Y_{w,j} = \sum_{j=1}^n h_{ij} Y_{w,j}, \end{aligned} \tag{2.12}$$

where

$$h_{ij} = \frac{1}{n} + \frac{(z_i - \bar{z})(z_j - \bar{z})}{Szz}.$$

So, the fitted value $\hat{Y}_{w,i}$ will be a weighted average of all the responses $Y_{w,j}$ with weights h_{ij}. Then, the weight exerted by an individual in its own prediction would be

$$h_{ii} = \frac{1}{n} + \frac{(z_i - \bar{z})^2}{Szz},$$

which will be large if its abscissa is far from the mean. The amount h_{ii} is known as dataset of the i-th data. It depends only on the value of explanatory variable, in this case z_i, and indicates the weight that $Y_{w,i}$ in his own fitting, $\hat{Y}_{w,i}$. Therefore it is interpreted as the attractiveness of the individual over the fitted line. Note that h_{ij} are elements of an idempotent and simetric matrix H. As a consequence $h_{ii} = \sum_{j=1}^{n} h_{ij}^2$.

To detect spatial outliers, regression residuals will be used. Note that the residuals may be written as

$$\begin{aligned}\hat{\varepsilon}_i &= Y_{w,i} - \hat{Y}_{w,i} = \beta_0 + \beta_1 z_i + \varepsilon_i - \sum_{j=1}^{n} h_{ij} Y_{w,j} = \beta_0 + \beta_1 z_i + \varepsilon_i - \sum_{j=1}^{n} h_{ij}(\beta_0 + \beta_1 z_j + \varepsilon_j) \\ &= \beta_0 + \beta_1 z_i + \varepsilon_i - \beta_0 \sum_{j=1}^{n} h_{ij} - \beta_1 \sum_{j=1}^{n} h_{ij} z_j - \sum_{j=1}^{n} h_{ij}\varepsilon_j = \varepsilon_i - \sum_{j=1}^{n} h_{ij}\varepsilon_j.\end{aligned}$$

The last equality is obtained by considering the following two expressions.

$$\sum_{j=1}^{n} h_{ij} = \sum_{j=1}^{n} \left[\frac{1}{n} + \frac{(z_i - \bar{z})(z_j - \bar{z})}{Szz}\right] = \frac{n}{n} + \frac{(z_i - \bar{z})}{Szz} \sum_{j=1}^{n} (z_j - \bar{z}) = 1.$$

$$\begin{aligned}\sum_{j=1}^{n} h_{ij} z_j &= \sum_{j=1}^{n} \left[\frac{1}{n} + \frac{(z_i - \bar{z})(z_j - \bar{z})}{Szz}\right] z_j = \bar{z} + \frac{(z_i - \bar{z})}{Szz} \sum_{j=1}^{n} (z_j - \bar{z}) z_j \\ &= \bar{z} + \frac{(z_i - \bar{z})}{Szz} \sum_{j=1}^{n} (z_j - \bar{z})^2 - \bar{z} \frac{(z_i - \bar{z})}{Szz} \sum_{j=1}^{n} (z_j - \bar{z}) = z_i.\end{aligned}$$

Under the assumption made by Anselin (1995), let ε_i be independent random variables, zero-mean with common variance σ^2. Using expectation and variance properties we find that

$$\mathbb{E}(\hat{\varepsilon}_i) = \mathbb{E}\left(\varepsilon_i - \sum_{j=1}^{n} h_{ij}\varepsilon_j\right) = \mathbb{E}(\varepsilon_i) - \sum_{j=1}^{n} h_{ij}\mathbb{E}(\varepsilon_j) = 0.$$

Moreover, following Anselin (1995),

$$\begin{aligned}\mathrm{Var}(\hat{\varepsilon}_i) &= \mathrm{Var}\left(\varepsilon_i - \sum_{j=1}^{n} h_{ij}\varepsilon_j\right) = \mathrm{Var}(\varepsilon_i) + \mathrm{Var}\left(\sum_{j=1}^{n} h_{ij}\varepsilon_j\right) - 2\mathrm{Cov}\left(\varepsilon_i, \sum_{j=1}^{n} h_{ij}\varepsilon_j\right) \\ &= \sigma^2 + \sum_{j=1}^{n} h_{ij}^2 \sigma^2 - 2h_{ii}\sigma^2 = \sigma^2(1 - h_{ii}).\end{aligned}$$

Since each residual has different variance, depending on its dataset, standardized residuals will be used to detect outliers. Given that the error variance is often unknown, it must be estimated. A natural estimator would be the sample variance, which will be denoted by $\hat{\sigma}^2$. Consequently, these new regression residuals are defined as

$$d_i = \frac{\hat{\varepsilon}_i}{\hat{\sigma}\sqrt{1 - h_{ii}}}.$$

Thus, observations with too large standardized residuals (in absolute value) show that the data is outlier in some way. It can be considered candidate to be an outlier the observations which have standardized residuals greater or less than 2 or -2, respectively, which would be approximately the quantiles of a standard normal containing more of the 95 % of the observations.

2.2.2. Variogram cloud

From a geostatistical perspective, the variogram cloud will be used as a graphical tool to detect spatial outliers. The variogram cloud, introduced by Cressie (1993), shows pairs of points attached by neighborly relations, providing an estimate of the dependence structure of the spatial process. That is, it would be an estimate of the basic geostatistical tool, the variogram, which was defined previously as follows

$$2\gamma(\mathbf{u}) = \mathrm{Var}(Z(\mathbf{s}+\mathbf{u}) - Z(\mathbf{s})), \quad \forall \mathbf{u} \in D. \tag{2.13}$$

Since the dependence structure is often unknown, (2.13) must be estimated. In this case, the estimator proposed by Cressie and Hawkins (1980) is considered:

$$2\bar{\gamma}(\mathbf{u}) = \frac{1}{(0.457 + 0.494/|N(\mathbf{u})|)} \left\{ \frac{1}{|N(\mathbf{u})|} \sum_{(i,j)\in N(\mathbf{u})} |Z(\mathbf{s}_i) - Z(\mathbf{s}_j)|^{1/2} \right\}^4, \tag{2.14}$$

where $|N(\mathbf{u})|$ is the number of pairs in $N(\mathbf{u}) \equiv \{(i,j) : \mathbf{s}_i - \mathbf{s}_j = \mathbf{u}\}$. If the process is isotropic), then the estimator of the variogram may be written as a function of $\|\mathbf{u}\| = u$. Therefore, in the variogram cloud, for each distance u, the value of the estimation of the variogram given in (2.14) is plotted. Small distances but with large values of the semivariogram may indicate a spatial outlier, although values in both locations do not represent outliers if the spatial dimension is omitted.

Variogram cloud requires non-trivial post-processing of highlighted pairs to separate spatial outliers from their neighbors, particularly when multiple outliers are present.

2.3. Some simulated examples

A simulation is performed to illustrate indicators and exploratory techniques introduced in the previous sections. Firstly, a sample is simulated regularly from a Gaussian spatial process (its finite-dimensional distribution is normal or Gaussian) considering the unit square as support and taking the exponential model given in (1.13) as covariance function. Based on this simulation scenario, different parameter values of the range and the sill in the covariance function (1.13) are considered, to see their effect regarding the presence of spatial outliers. Finally, an alteration of an observation is realized: a non spatial attribute excessively high for the simulated sample is chosen, so this observation would be an outlier.

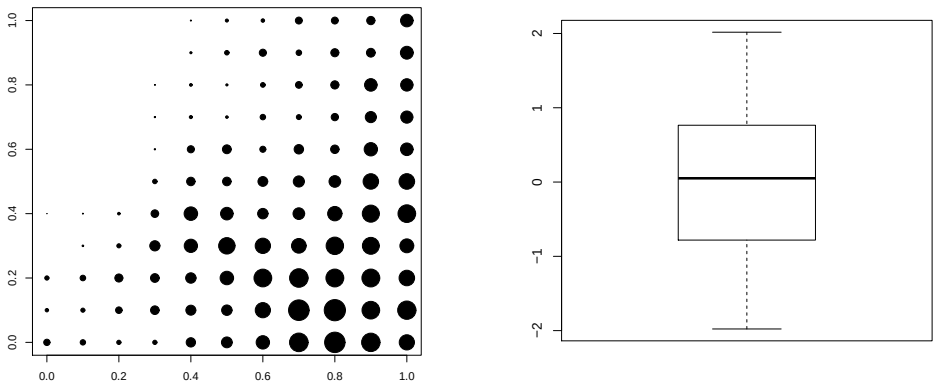

Figure 2.1: Scatterplot (right) and box-plot (right) omitting the spatial dimension. Sample size: 121. Regularly simulated data from a Gaussian spatial process on unit square with exponential covariogram (1.13) with parameters $c_0 = 0$, $c_e = 1$ and $a_e = 2$.

With a sample size of 121, considering the unit square as support, data have been drawn taking the exponential model given in (1.13) as covariance function with $c_0 = 0$, $c_e = 1$ and $a_e = 2$. Figure 2.1 (left) plots the sample data points of the spatial process simulated with symbol area proportional to observation measure. Observations will be numbered from 1 to 121, starting from the bottom-left corner, and advancing by rows until the upper-right corner in the unit square. It is worth noting that realizations of a process with constant mean, but strong spatial correlation, frequently seem to present trends, as it can be seen in Figure 2.1 (left). In this case, the range value, 2, is greater than any of the distances in the unit square, leading a strong dependence. The box-plot of the attribute values (see Figure 2.1, right) shows that there is no global outliers (omitting the spatial dimension). Nevertheless, to detect spatial outliers it is necessary to go beyond.

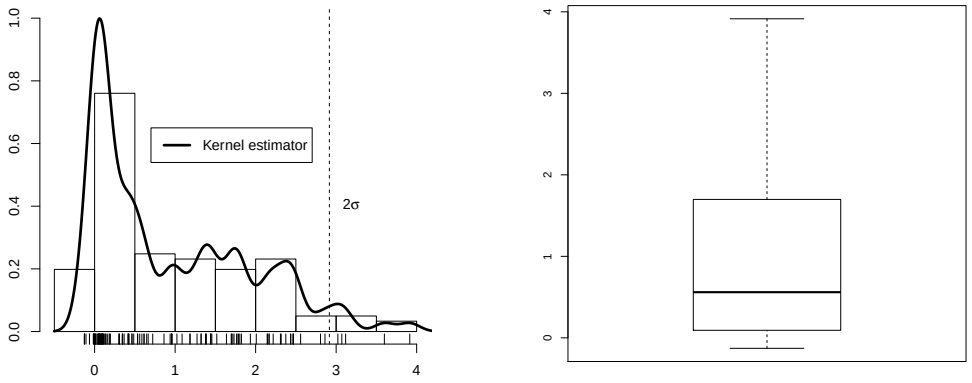

Figure 2.2: Histogram (left) with Gaussian kernel density estimator and boxplot (right) of I_i. Sample size: 121. Regularly simulated data from a Gaussian spatial process on unit square with exponential covariogram (1.13) with parameters $c_0 = 0$, $c_e = 1$ and $a_e = 2$.

When the nearest neighbor ($k = 1$) is used to compute the weight matrix, Moran's I coefficient is 0.9495, larger than $-1/(n-1)$, which indicates positive spatial autocorrelation. Similar conclusions are obtained from Geary's c, which value is 0.075, being close to zero (maximal positive autocorrelation). Regarding local indicators, as it was mentioned previously, the distribution of I_i statistics for the sample can provide an indication of outliers by means of a 2σ rule. The mean of the distribution of the I_i is Moran's I, or 0.9495, and twice the standard deviation from the mean to the value of 2.9159. This is exceeded by the observation 9, with a value of 3.1178 for I_i, by the observation 10, with a value of 3.0703 for I_i, by the observation 20, with a value of 3.9145 for I_i, by the observation 21, with a value of 3.0214 for I_i and by the observation 91, with a value of 3.5989 for I_i. Although it is not a test, it provides useful insight into the nature of these five observations. Four of them $(9, 10, 20$ and $21)$ located in the bottom right corner of Figure 2.1 (left).

Figure 2.2 (left) plots the histogram of I_i, a non-parametric estimator of the density function, with Gaussian kernel density estimator (the method of Sheather and Jones to select the bandwidth is used). There are five observations which exceed the 2σ threshold. Figure 2.2 (right) shows the boxplot of I_i, which indicates that there is not any outlier in the sample of I_i.

Figure 2.3 (left) plots the variogram cloud for these simulated data. There are few pairs of data for which have small distances and large semivariogram values. There are not observations candidates to be spatial outliers. Figure 2.3 (right) shows the Moran scatterplot for this simulate data with the regression line that best fits the data by least squares. In this case, the single nearest neighbour ($k = 1$) will be used to compute the weight matrix, W. It can be seen that many of the observations fall into

the lower left and upper right quadrants (109 out of 121), showing a positive association. Since the pairs are given for standardized values, outliers may be easily visualized as points further than two units away from the origin. The observation 20 has the value for the attribute that is higher than two standard deviations from mean (on the vertical axis of the Figure 2.3, right), while the observation 21 also has values for the spatial lag that are twice the mean (horizontal axis of the Figure 2.3, right). Observations $10, 18, 69, 80$ and 91 could be outliers too, since they are almost two deviations of the origin.

If standarized regression residuals of the linear regression made in the corresponding Moran Scatterplot are considered, then there are ten observations which have standarized residuals greater or less than 2 or -2, respectively.

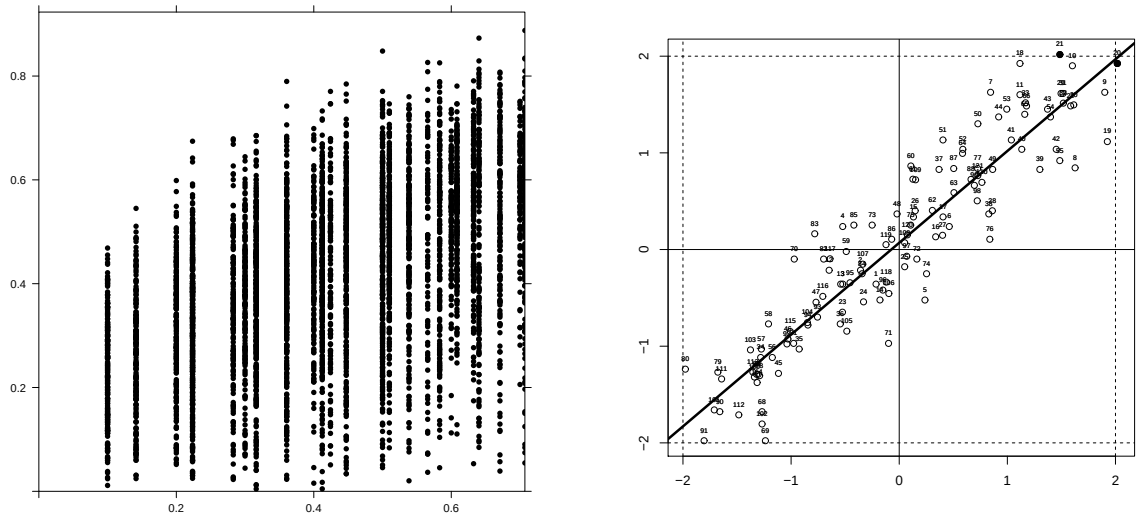

Figure 2.3: Exploratory tools for detecting spatial outliers: variogram cloud (left) and Moran scatterplot with $k = 1$ (right). Sample size: 121. Regularly simulated data from a Gaussian spatial process on unit square with exponential covariogram (1.13) and parameters $c_0 = 0$, $c_e = 1$ and $a_e = 2$. Outliers in the Moran scatterplot, solid points.

Other values of k have been also considered to check the effect of this tuning parameter. When $k = 60$ is considered to compute the matrix of weights, the spatial association is still positive, although lower, and the mean of I_i or Moran's I is 0.4702. Furthermore, twice the standard deviation from the mean to the value is 1.3541. This is only exceeded by the observation 20, with a value of 1.4684 for I_i. Figure 2.4 (left) presents the Moran scatterplot considering $k = 60$. The observation 20 has the value for the attribute that is are more than two standard deviations higher than the mean (on the vertical axis of the Figure 2.4 (left)), then it may be an outlier. Observations $9, 19$ and 80 could be outliers too, since they are almost two deviations of the origin. If standarized regression residuals of the linear regression made in the corresponding Moran Scatterplot are considered, then there are five observations which have standarized residuals greater or less than 2 or -2, respectively.

When $k = 120$ is considered to compute the matrix of weights, there are not spatial association, since the mean of the distribution of the I_i or Moran's I is -0.0083 (equal to $-1/(n-1)$). Furthermore, twice the standard deviation from the mean to the value is -0.0252 or 0.0085, which is exceeded by the observation 9, with a value of -0.0304 for I_i, by the observation 19, with a value of -0.0311 for I_i, by the observation 20, with a value of -0.0342 for I_i, by the observation 80, with a value of -0.0329 for I_i and by the observation 91, with a value of -0.0274 for I_i. These five observations deserve closer scrutiny. Figure 2.4 (right) shows the Moran scatterplot considering $k = 120$. The observation 20 has

the value for the attribute that is are more than two standard deviations higher than the mean (on the vertical axis of the Figure 2.4 (right)), then it may be an outlier. Observations 9, 19 and 80 could be outliers too, since they are almost two deviations of the origin. If standarized regression residuals of the linear regression obtained from the corresponding Moran Scatterplot are considered, then there are two observations which have standarized residuals greater or less than 2 or -2, respectively.

In this case the choice of k varies the possible candidates to be outliers. The observation 20 could be an spatial outlier, but in all cases, this observation is in the limit of the two standard deviations higher than the mean in the Moran scatterplot. However, from variogram cloud does not seem to have outliers.

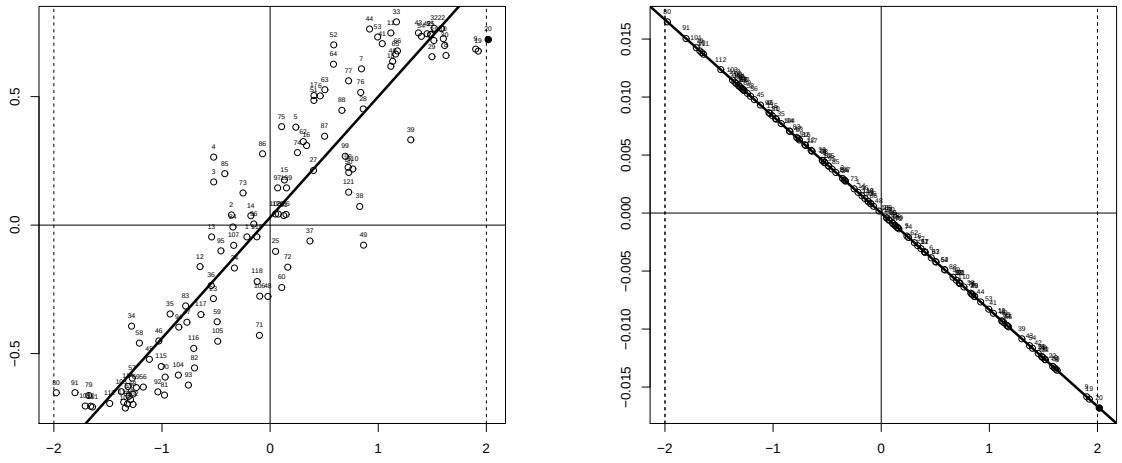

Figure 2.4: Moran scatterplots with $k = 60$ (left) and $k = 120$ (right). Sample size: 121. Regularly simulated data from a Gaussian spatial process on unit square with exponential covariogram (1.13) with parameters $c_0 = 0$, $c_e = 1$ and $a_e = 2$. Outliers in the Moran scatterplot, solid points.

2.3.1. Effect of the range parameter

Under the same simulation scenario (121 regularly spaced Gaussian data in the unit square with exponential covariance), different parameter values in the covariance function (1.13) are considered, to see their effect regarding the presence of outliers. Now, a sample of size 121 is simulated regularly from a Gaussian spatial process considering the unit square as support, taking $c_0 = 0$, $c_e = 1$ and $a_e = 0.1$ as parameters in the covariance function (1.13). In this case the range is smaller than in the previous simulation, resulting in a weaker dependence. So, the effect of the range parameter will be reflected. The single nearest neighbour ($k = 1$) will be used to compute the weight matrix, W.

Moran's I coefficient is 0.6417, larger than $-1/(n-1)$, which indicates positive spatial autocorrelation. Similar conclusions are obtained from Geary's c, which value is 0.4781, being close to zero (maximal positive autocorrelation). The autocorrelation is smaller than in the previous simulation. Moreover, the threshold value is 3.1641. This is exceeded by nine observations, however, varying the value of k, the observations which exceed twice the standard deviation from the mean to the value corresponding are different. For example, if $k = 120$ is considered, then observations 8, 19, 34 and 80 exceed the 2σ threshold.

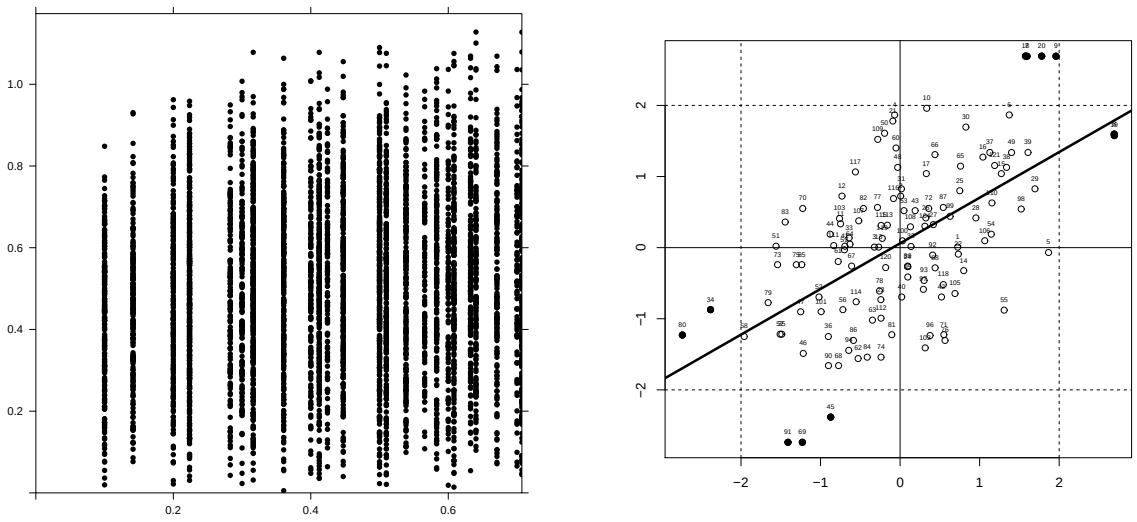

Figure 2.5: Exploratory tools for detecting spatial outliers: variogram cloud (left) and Moran scatterplot with $k = 1$ (right). Sample size: 121. Regularly simulated data from a Gaussian spatial process with exponential covariogram (1.13) with parameters $c_0 = 0$, $c_e = 1$ and $a_e = 0.1$. Outliers in the Moran scatterplot, solid points.

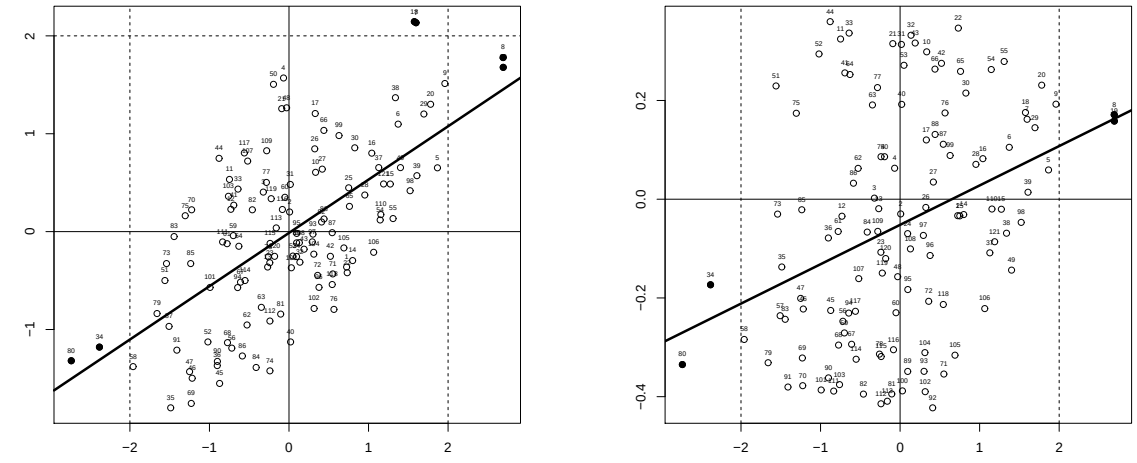

Figure 2.6: Moran scatterplots with $k = 2$ (left) and $k = 70$ (right). Sample size: 121. Regularly simulated data from a Gaussian spatial process on unit square with exponential covariogram (1.13) with parameters $c_0 = 0$, $c_e = 1$ and $a_e = 0.1$. Outliers in the Moran scatterplot, solid points.

Figure 2.5 (left) plots the variogram cloud of this data. It can be seen that there are many pairs of data which represent small distances and large semivariogram values (larger than when $a_e = 2$ was considered in the exponential covariogram (1.13)). Observations $8, 19, 49, 51$ and 80 can be identified as spatial outliers since they appears in the data pairs mentioned. Figure 2.5 (right) is the corresponding

Moran scatterplot with $k = 1$. There are many observations which are more than two units of the origin. However, these observations are not spatial outliers, because, according to the choice of k, the conclusions are not the same. Figure 2.6 (left) presents the Moran scatterplot of this data (with $k = 2$), as it can be seen that observations $9, 20, 45, 69$ and 91 are not candidates to be outliers. If $k = 80$ is considered, observations 7 and 18 are not too (see Figure 2.6, right). According to the Moran scatterplot, the observations $8, 19, 34$ and 80 are candidates to be outliers.

If standarized regression residuals of the linear regression made in the corresponding Moran Scatterplot are considered, then there are four, two and three observations (considering $k = 1, 2, 70$, respectively) which have standarized residuals greater or less than 2 or -2, respectively.

In this case, spatial outliers are influenced by the choice of the range parameter.

2.3.2. Effect of the sill parameter

Now, a sample of size 121 is simulated regularly from a Gaussian spatial process considering the unit square as support, and taking $c_0 = 0$, $c_e = 0.1$ and $a_e = 2$ as parameters in the covariance function (1.13). So, the effect of the sill parameter will be reflected (a value smaller of the sill is considered now, which was $c_e = 1$ before). The single nearest neighbour ($k = 1$) will be used to compute the weight matrix, W.

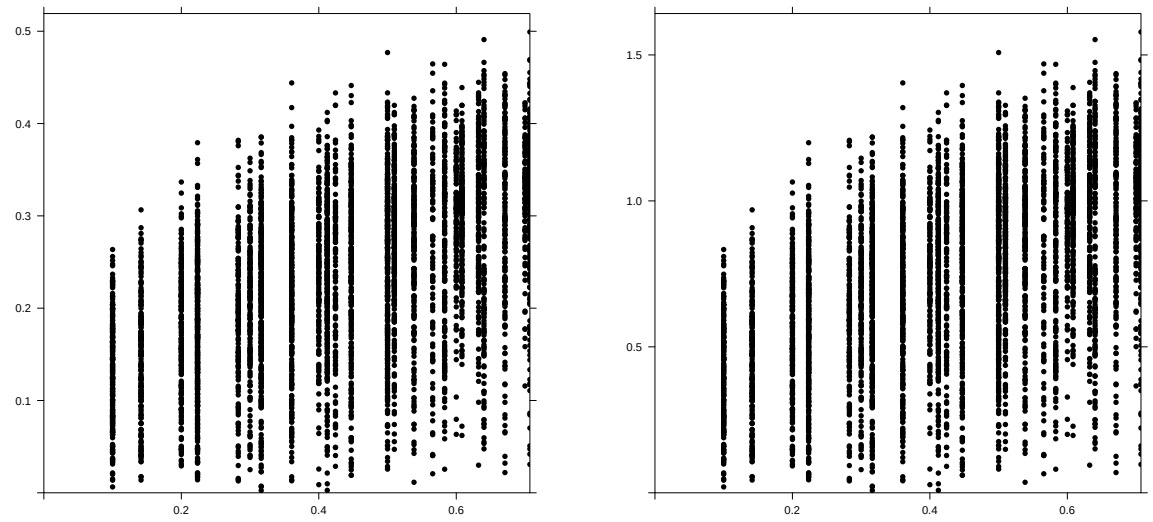

Figure 2.7: Exploratory tools for detecting spatial outliers: variogram clouds. Sample size: 121. Regularly simulated data from a Gaussian spatial process on unit square with exponential covariogram (1.13) with parameters $c_0 = 0$, $c_e = 0.1$ and $a_e = 2$ (left). Sample size: 121. Regularly simulated data from a Gaussian spatial process on unit square with exponential covariogram (1.13) with parameters $c_0 = 0$, $c_e = 10$ and $a_e = 2$ (right).

Moran's I coefficient is 0.9495, larger than $-1/(n-1)$, which indicates positive spatial autocorrelation. Similar conclusions are obtained from Geary's c, which value is 0.0746, being close to zero (maximal positive autocorrelation). This values are equal than in the first simulation (local Moran statistics are equal too). So, same conclusions about spatial outliers are obtained that when the first simulation was performed (remember that the sill is the covariance at zero distance).

Moran scatterplot is exactly the same than in Figure 2.3 (right). Variogram cloud for these data is plotted in Figure 2.7 (left). This representation differs from the Figure 2.3 (left) in that the range of values of the semivariogram is less comprehensive, but the conclusions about spatial outliers are

identical. If standarized regression residuals of the linear regression made in the corresponding Moran Scatterplot are considered, then there are ten observations (the same as in the first simulation done) which have standarized residuals greater or less than 2 or -2, respectively.

Now, a sample of size 121 is simulated regularly from a Gaussian spatial process considering the unit square as support, and taking $c_0 = 0$, $c_e = 10$ and $a_e = 2$ as parameters in the covariance function (1.13). A value larger than the sill is considered now.

Moran scatterplot is exactly the same than when other values of the sill were considered, hence the possible outliers would be the observation 20 . Figure 2.7 (right) plots the variogram cloud for this simulated data. In this case, for all distances, values of he semivariogram are larger than when $c_e = 0.1$ and $c_e = 1$ were considered. However, same conclusions about spatial outliers are obtained (see Figure 2.3 and Figure 2.7 (left) and (right) for comparison). If standarized regression residuals of the linear regression made in the corresponding Moran Scatterplot are considered, then the results obtained are identical to when different values of c_e are chosen. Therefore, the sill parameter does not seem to affect the possible outliers.

2.3.3. Alteration of an observation

A sample of size 121 is simulated regularly from a Gaussian spatial process considering the unit square as support and taking the exponential model given in (1.13) as covariance function with $c_0 = 0$, $c_e = 1$ and $a_e = 2$. The observation 61 (central square) is modified with the value of 5.1365 (median+$3IQR$, where IQR denotes the interquartile range of the observed values of the process): a non spatial attribute excessively high for the simulated sample, so this would be an outlier. The single nearest neighbour ($k = 1$) will be used to compute the weight matrix, W.

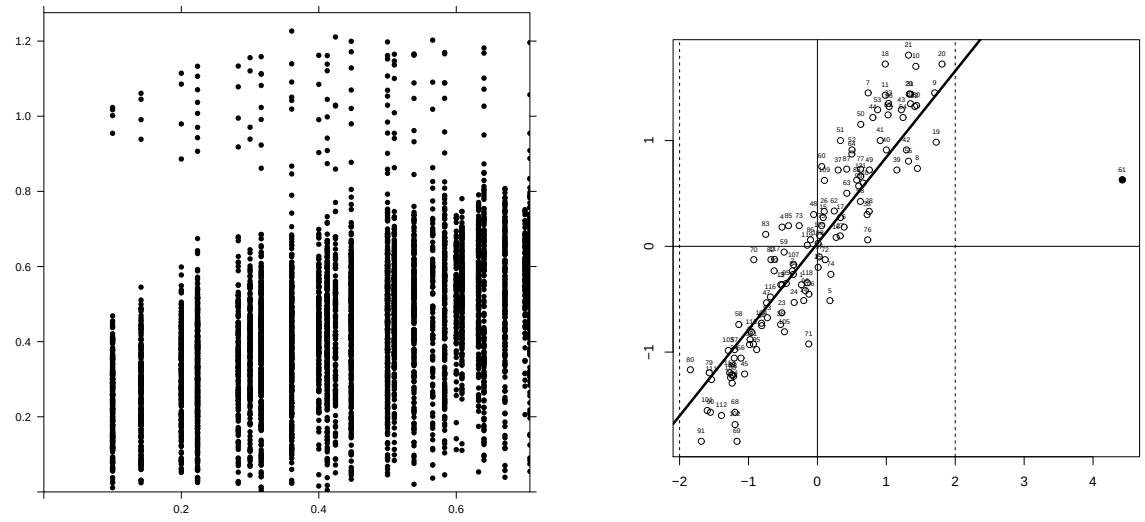

Figure 2.8: Exploratory tools for detecting spatial outliers: variogram cloud (left) and Moran scatterplot with $k = 1$ (right). Sample size: 121. Regularly simulated data from a Gaussian spatial process on unit square with exponential covariogram (1.13) with parameters $c_0 = 0$, $c_e = 1$ and $a_e = 2$.

Moran's I coefficient is 0.8248, larger than $-1/(n-1)$, which indicates positive spatial autocorrelation. Similar conclusions are obtained from Geary's c, which value is 0.1167, being close to zero (maximal positive autocorrelation). The mean of the distribution of the I_i is Moran's I, or 0.8248, and twice the standard deviation from the mean to the value of 2.5161. This is exceeded by the observation 9, with a value of 2.5263 for I_i, by the observation 20, with a value of 3.1872 for I_i, by the observation

61, with a value of 2.7373 for I_i, by the observation 91, with a value of 3.1687 for I_i and by the observation 101, with a value of 2.5311 for I_i.

Figure 2.8 (left) shows the variogram cloud for these simulated data. It can be seen that there are many pairs of data which represent small distances and large semivariogram values. The observation 61 can be identified as a spatial outlier since it appears in all data pairs mentioned. Figure 2.8 (right) is the Moran scatterplot for this simulate data with the regression line that best fits the data by least squares. Since the pairs are given for standardized values, outliers may be easily visualized as points further than two units away from the origin. The central observation has the value for the attribute that is are more than two standard deviations higher than the mean (on the vertical axis of the Figure 2.8 (right)), then it is candidate to be a spatial outlier.

If standarized regression residuals of the linear regression made in the corresponding Moran Scatterplot are considered, then there are two observations (61 and 69) which have standarized residuals greater or less than 2 or -2, respectively.

If the value of k-nearest neighbors is larger, then the number of observations which exceed twice the standard deviation from the mean to the value corresponding are smaller than five. Moreover, in the Moran scatterplot the only observation which would still have the value for the attribute that is are more than two standard deviations higher than the mean is the central observation. Therefore, the central observation is candidate to be outlier. To emphasize that, from a boxplot this observation could be identified as a global outlier, thus omitting the spatial dimension. If a smaller value for a_e in the expression (1.13) is considered, both tools would detect the central observation as atypical, as long as this value of a_e is greater than 0.8. For lower values, exploratory techniques could identify as outliers other observations too.

2.4. A final comment on exploratory tools

Once again, a sample of size 121 is simulated regularly from a Gaussian spatial process considering the unit square as support and taking the exponential model given in (1.13) as covariance function with $c_0 = 0$, $c_e = 1$ and $a_e = 0.15$. The observation 61 (central square) is modified with the value of 4.4232 (median+3IQR, where IQR denotes the interquartile range).

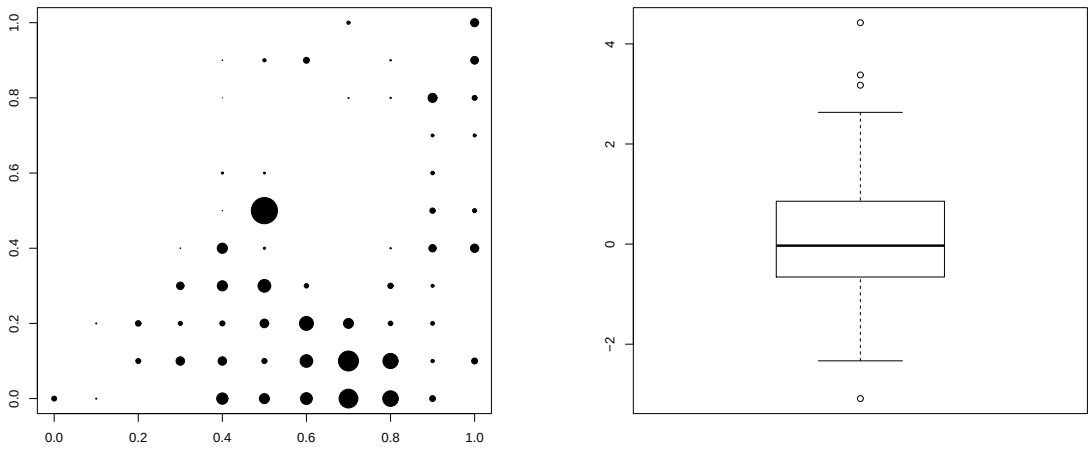

Figure 2.9: Symbol area proportional to atributtes concentration (left) and Box-plot (right) omitting the spatial dimension. Sample size: 121. Regularly simulated data from a Gaussian spatial process on unit square with exponential covariogram (1.13) and parameters $c_0 = 0$, $c_e = 1$ and $a_e = 0.15$.

Figure 2.9 (left) plots the sample data points of the spatial process simulated with symbol area proportional to measured. The box-plot of the attribute values (see Figure 2.9, right) shows that there are four global outliers, $8, 19, 61$ and 80 (omitting the spatial dimension).

Figure 2.10 (left) plots the variogram cloud for these simulated data. It can be seen that there are many pairs of data which represent small distances and large semivariogram values. The observation 61 can be identified as spatial outliers since in appears in all data pairs mentioned. In this case, there is just an outlier. However, when multiple outliers are present, it is difficult detect outliers in the variogram cloud, as there may be many points which are at small distances and have large variogram value. Some algorithm of exhaustive search should be used.

Figure 2.10 (right, full line) shows the number of spatial outliers (observations further than two units away from the origin) in Moran Scatterplot, for different values of k. The number of outliers decreases with increasing k, stabilizing to 5 from $k = 4$ (it is always the same five candidate data). Figure 2.10 (right, dashed line) presents the number of spatial outliers for different values of k too, based on standarized regression residuals of the linear regression realized in the Moran scatterplot. The behaviour is not monotonous, and the possibles candidates to be outliers are different in function of the value of k. This criterion gives less outliers than the others. Figure 2.10 (right, dotted line) plots the number of spatial outliers for different values of k, using the 2σ threshold for the distribution of I_i. Again, the behaviour is not even monotone, in fact, a stable pattern is not obtained varying the value of k. However, when the same number of outliers are obtained, the same data are candidates to be outliers.

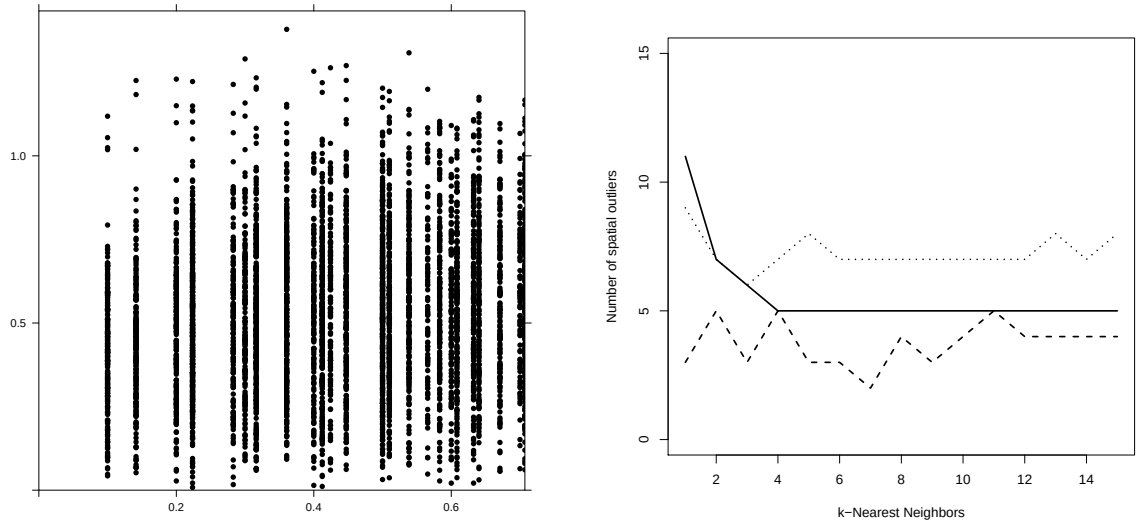

Figure 2.10: Variogram cloud (left). Number of spatial outliers in function of k in Moran Scatterplot (right, full line), using standarized regression residuals (right, dashed line) and using the 2σ threshold (right, dotted line). Sample size: 121. Regularly simulated data from a Gaussian spatial process on unit square with exponential covariogram (1.13) and parameters $c_0 = 0$, $c_e = 1$ and $a_e = 0.15$.

In this case the choice of k varies the possible candidates to be outliers. Moreover different methods are inconsistent in their conclusion. They propose different candidates to be outliers, considering the same value of k.

We should also notice that there is a possibility too strong simplification in Anselin (1995) procedure. In the construction of Moran scatterplot described in the Section 2.2.1, the error term in 2.11 is considered as independent. However, by the constructing of $Y_{w,i}, i = 1, \ldots, n$, this is not true. The covariance matrix of $\mathbf{Y}_w$ (being this vector the one that collects all the $Y_{w,i}$ values) is given by

$\text{Cov}(\mathbf{Y}_w, \mathbf{Y}_w) = W\Sigma$ (where Σ is the original covariance matrix and W is the weight matrix), which is not a diagonal matrix. It would be interesting to explore what happens with Moran scatterplot if outliers are identified according for this dependence, which would require a pilot estimator of Σ.

The problem of detection of outliers in spatial data is not solved given that existing exploratory techniques do not allow a correct identification of such data (even if it is certain that its definition is not precise). However, the consideration of outliers in inferential procedures may provide questionable results, therefore it is necessary the use of techniques that mitigate their effect. In the next chapter, a proposal for the spatial trend estimation (when there are spatial outliers) will be carried out.

2.5. Illustration with real data

In the present illustration, we use measurements of mercury concentration[2] (Hg in ppb) in moss sample surveys collected in Galicia in July of 2000 (see Fernández et al (2005)). Sampling stations were located at the vertices of a 15×15 km UTM grid. A total of 148 sampling sites was distributed over the entire region and the adjacent border area. The principal moss species collected was Scleropodium purum and in its absence, Hypnum cupressiforme was collected. Samples were collected at least $300m$ from main roads and any populated areas and at least $100m$ from other kinds of roads and isolated houses. Where possible, samples were collected from open areas, otherwise, they were collected from clearings within forest areas.

Figure 2.11 (left) is the sample data point for mercury (ppb), with symbol area proportional to measured concentration. The box-plot of the attribute values (see Figure 2.11, right) shows that there is no global outliers (omitting the spatial dimension).

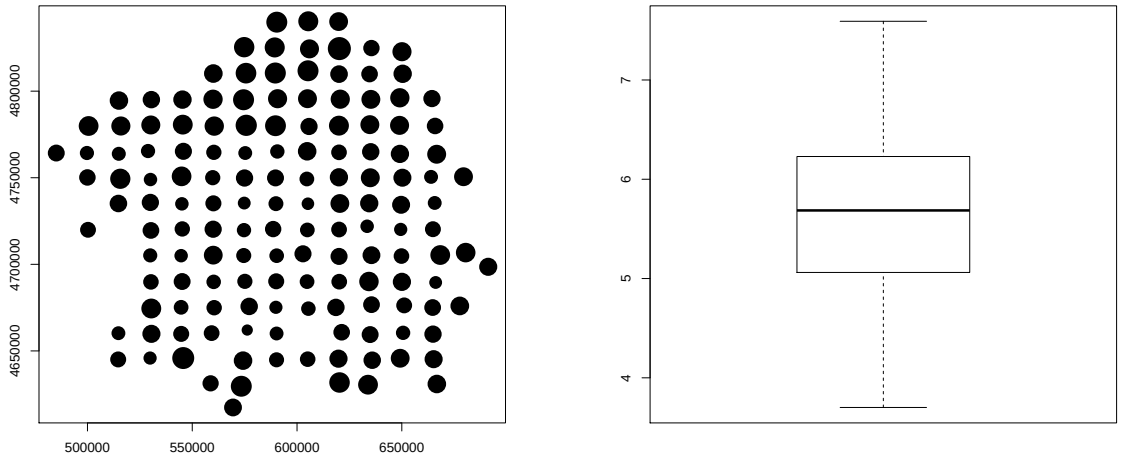

Figure 2.11: Symbol area proportional to atributtes concentration (left) and Box-plot (right) omitting the spatial dimension. Measurements of mercury concentration (Hg in ppb) in moss sample surveys collected in Galicia in July of 2000.

When the nearest neighbor is used to compute the weight matrix, Moran's I coefficient is 0.3756, larger than $-1/(n-1)$, which indicates positive spatial autocorrelation. Regarding local indicators, the distribution of I_i statistics for the sample may provide an indication of outliers by means of a 2σ

[2]The author acknowledge the group of Ecotoxicoloxía e Ecofisioloxía Vexetal of the University of Santiago de Compostela for providing the dataset used for illustrating the methods.

rule. The mean of the distribution of the I_i is Moran's I, or 0.3756, and twice the standard deviation from the mean to the value of 2.1057. This is exceeded by the observation 11, with a value of 2.5133 for I_i, by the observation 12, with a value of 2.5133 for I_i, by the observation 33, with a value of 2.7575 for I_i, by the observation 34, with a value of 2.7575 for I_i and by the observation 126, with a value of 3.5894 for I_i.

Figure 2.12 (left) plots the variogram cloud for these sample data. It can be seen that there are many pairs of data which represent small distances and large semivariogram values. However, only the observation 133 can be identified as spatial outliers since it appears in many data pairs mentioned.

Moran scatterplot for these concentrations of mercury with the regression line that best fits the data by least squares is plotted in the Figure 2.12 (right). In this case, the single nearest neighbour ($k = 1$) will be used to compute the weight matrix, W. It can be seen that many of the observations fall into the lower left and upper right quadrants (94 out of 148), showing a positive association. Since the pairs are given for standardized values, outliers may be easily visualized as points further than two units away from the origin. The observations $7, 125$ and 133 have values for the attribute that are higher than two standard deviations from mean (on the vertical axis of the Figure 2.12, right), while the observations 123 and 126 also have values for the spatial lag that are twice the mean (horizontal axis of the Figure 2.12, right).

If standarized regression residuals of the linear regression made in the corresponding Moran Scatterplot are considered, then there are four observations which have standarized residuals greater or less than 2 or -2, respectively. These observations would be $75, 122, 123$ and 126.

Possible common candidate for the methods that derive of Moran Scatterplot will be the observation 126. However, this conclusion is not true, as shown below.

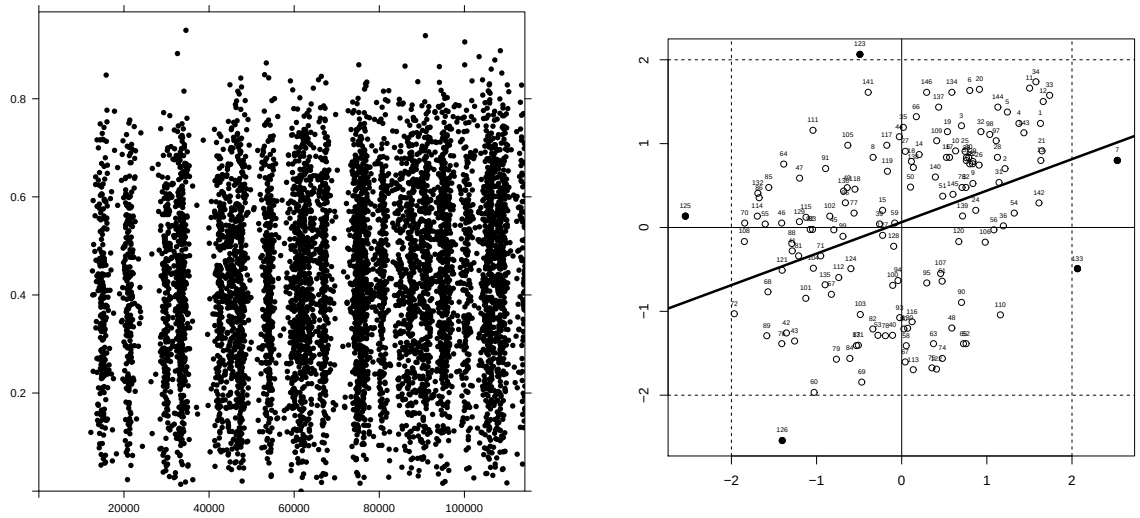

Figure 2.12: Exploratory tools for detecting spatial outliers: variogram cloud (left) and Moran scatterplot with $k = 1$ (right). Measurements of mercury concentration (Hg in ppb) in moss sample surveys collected in Galicia in July of 2000. Outliers in the Moran scatterplot, solid points.

Other values of k have been also considered to check the effect of this tuning parameter. When $k = 2$ is considered to compute the matrix of weights, the spatial association continue being positive, however this is lower, the mean of the distribution of the I_i or Moran's I is 0.3321. Furthermore, twice the standard deviation from the mean to the value is 1.8427. This threshold is exceeded by eleven observations. Figure 2.13 (left) presents the Moran scatterplot considering $k = 2$. The observations $7, 125$ and 133 have values for the attribute that are higher than two standard deviations from mean

(on the vertical axis of the Figure 2.13, left), while the observations 6, 113 and 126 also have values for the spatial lag that are twice the mean (horizontal axis of the Figure 2.13, left). Note that, observation 113 is a candidate to be an outlier for this value of k. This is due to the two nearest neighbors are observations 114 and 125 (which is also candidate to be an outlier), while, the nearest neighbor is the observation 114.

If standarized regression residuals of the linear regression made in the corresponding Moran Scatterplot are considered, then there are eight observations which have standarized residuals greater or less than 2 or -2, respectively.

In this case, increasing k, methods based on Moran Scatterplot detect more possible candidates to be outliers.

When $k = 147$ is considered to compute the matrix of weights, the spatial association continue being positive. Furthermore, twice the standard deviation from the mean to the value is 1.7343, which is exceeded by twelve observations. Figure 2.4 (right) shows the Moran scatterplot considering $k = 3$. The observations 7, 125 and 133 has the value for the attribute that is are more than two standard deviations higher than the mean (on the vertical axis of the Figure 2.13 (right)), then they may be outliers.

If standarized regression residuals of the linear regression made in the corresponding Moran Scatterplot are considered, then there are six observations which have standarized residuals greater or less than 2 or -2, respectively. These observations would be $6, 54, 55, 113, 133$ and 141.

In this case the choice of k varies the possible candidates to be outliers. Furthermore, it can be seen that the methods, including variogram cloud, are inconsistent in the sense that they detect different outliers, as we had mentioned in the Section 2.4.

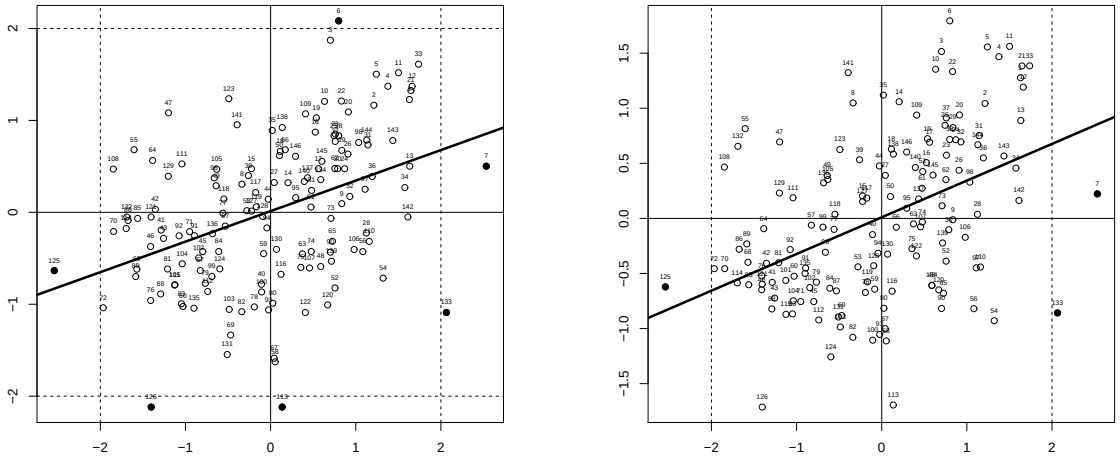

Figure 2.13: Moran scatterplots with $k = 2$ (left) and $k = 3$ (right).Measurements of mercury concentration (Hg in ppb) in moss sample surveys collected in Galicia in July of 2000. Outliers in the Moran scatterplot, solid points.

Chapter 3

Dealing with outliers in spatial regression

In geostatistics the observed data tend to exhibit an important feature: close observations tend to be more similar than observations which are far apart. Such observations cannot be treated as independent and the dependence structure should be taken into account in any descriptive or inferential procedure. In particular, from the perspective of regression models, for example through trend surface models, the dependence structure should be considered and properly introduced into the model.

Trend surface models try to relate observations of a spatial process which varies continuously with the geographic locations in which those observations are taken, trying to establish patterns of increase or decrease with respect to the coordinates through (usually linear) regression models. In these models, two sources of variability can be distinguished (see Cressie (1993)): a regression function or trend which would gather large-scale variability and the error term, which would represent small-scale variability. The main difference between these models and the classical linear regression is that the errors present a dependence structure, which is usually assumed to be intrinsically stationary or second-order. That dependence structure is usually unknown but should be included in the model, through the covariogram (if the process is second-order stationary) or the variogram (if the process is intrinsically stationary).

The problem of trend surface estimation can be solved through least squares tools (where pilot estimations of the variogram are used) or maximum likelihood (the estimation of the parameters of the trend and of the dependence are approached jointly, under the assumption of normality), as it is described in Diggle and Ribeiro (2007). Even so, the trend surface estimation has only been discussed in the literature in the case of "simple" statistical processes, under the assumption of normality (see Crujeiras and Van Keilegom (2010)). However, as it was mentioned in the Introduction, even under this distributional premise, observed samples may present certain complexities, as the presence of outliers.

The aim of this chapter is to propose a new procedure to perform the trend surface models estimation which protect the conclusions against the influence of outliers. The idea would be to combine the trend surface models estimation using iterative least squares (taking into account the dependence structure) with the use of pseudo-data (obtained by a previous smoothing procedure) to mitigate the effect of outliers (see Akritas (1996) or Cristobal et al. (1987)).

In this chapter, the trend surface models estimation is reviewed, as well as the use of "pseudo-data" in order to mitigate the effect of outliers in regression. We will also propose a procedure of trend surface models estimation with outliers. A simulation study is also performed, as well as illustration with real data. Some discussion on the possible flaws of the method and issues for further study will be also mentioned.

3.1. Trend surface models

A widely used class of regression models in the spatial setting is the so called trend surface models (see Zimmerman and Stein (2010)), where the trend is considered as a polynomial function of the geographic coordinates (planar and quadratic, for instance).

Consider a random spatial process $\{Z(\mathbf{s}), \mathbf{s} \in D\}$. Assume that the behavior of the process is described by the following spatial regression model:

$$Z(\mathbf{s}) = X(\mathbf{s})'\beta + \varepsilon(\mathbf{s}), \quad \mathbf{s} \in D, \quad D \subset \mathbb{R}^d, \tag{3.1}$$

where $X(\cdot) \in \mathbb{R}^q$ denote the spatial regressors that may be geographic coordinates and $m(\cdot) = X(\cdot)'\boldsymbol{\beta}$ is a linear trend component (the superscript $'$ denotes the transpose), which captures the large-scale variability of the process. $\varepsilon(\cdot)$ denotes the random component and shows the local behavior or small-scale evolution. The error process $\varepsilon(\cdot)$ is assumed to be zero-mean and second-order stationary.

When the mean function $m(\cdot)$ is taken to be just a polynomial function of the geographic coordinates (as in the regression model given in (3.1)), then such models are called trend surface models. For example, the first -order (planar) and second -order (quadratic) polynomial trend surface models for the mean of a two-dimensional process are respectively as follows, where $\mathbf{s} = (s_1, s_2)$:

$$m(\mathbf{s}) = \beta_0 + \beta_1 s_1 + \beta_2 s_2, \tag{3.2}$$

$$m(\mathbf{s}) = \beta_0 + \beta_1 s_1 + \beta_2 s_2 + \beta_{11} s_1^2 + \beta_{12} s_1 s_2 + \beta_{22} s_2^2. \tag{3.3}$$

Using a full qth-order polynomial (a polynomial that includes all pure and mixed monomials of degree less than q) is recommended because this will ensure that the fitted model is invariant to the choice of origin and orientation of the coordinate system.

3.1.1. Estimation

Consider a random spatial process $\{Z(\mathbf{s}), \mathbf{s} \in D\}$ and suppose that its behaviour is described by the spatial regression model given in (3.1). Assume that $\varepsilon(\cdot)$ is a zero-mean and second-order stationary process.

For modelling the behaviour of the process, as well as for model-based spatial prediction, it is needed to know the structure of the process, that is, we must have an estimation of the components. Therefore, two sets of parameters must be estimated: $\boldsymbol{\beta} \in \mathbb{R}^q$, from the trend part and $\boldsymbol{\theta}$, the parameters from the dependence structure. Usually, $\boldsymbol{\theta}$ includes the range, the point variance (sill) and possibly the nugget effect (microscale variation).

With the decomposition given in (3.1), it is not clear which structure (first order or second order) must be estimated first. In practice, iterative procedures are a solution: estimate trend, compute residuals, estimate dependence and adjust trend. Maximum Likelihood methods overcome this problem, at the cost of specifying a full parametric model.

In trend surface models, the method of iteratively least squares is often used to approximate the model coefficients (see Diggle and Ribeiro (2007)). Nevertheless, this procedure is not very robust in the presence of outliers in the response.

Crujeiras and Van Keilegom (2010) proposed an estimation procedure which is a generalization of the method suggested by Gallant and Goebel (1976) in the context of temporally autocorrelated errors. Their purpose is to adapt the method suggested by Gallant and Goebel (1976) to a more general setting, for a spatial regression model (non-linear, our case will be linear) with spatially dependent errors, but without imposing any structural assumption.

The estimation procedure consists on three steps: (1) unweighted least squares estimation of $\boldsymbol{\beta}$, ignoring the dependence structure of the errors; (2) estimation of the variance-covariance matrix of the errors based on the estimator of $\boldsymbol{\beta}$ found in the first step; (3) weighted least squares estimation of $\boldsymbol{\beta}$, taking the dependence structure of the errors into account.

3.2. Outliers in regression. The use of "pseudo-data"

The idea of use "pseudo-data" in regression to mitigate the effect of certain complexities in data has been suggested by Akritas (1996) and Cristobal et al. (1987). In this last reference, with the goal of robustifying the estimation procedure, while Akritas (1996) proposes a procedure which works when incomplete data occur due to censoring and truncation.

Both procedures basically consists in replacing the observed responses by a nonparametric estimator. Let Y be the variable of interest, and let X denote an observable covariate. It will be assumed that X is univariate. The idea proposed by Akritas (1996) is to associate with each observed covariate value x_i an estimate $\hat{m}(x_i)$ of location of the conditional distribution of the response variable Y given $X = X_i$. For each covariate value x_i, $\hat{m}(x_i) = T(\hat{F}_i)$, where $\hat{F}_i$ is an estimator of the distribution (product-limit estimator for the cases considered in Akritas (1996)), evaluated from the response variable of the data points that belong to the window around X_i, and T is a functional (similar to a trimmed mean) specified by (3.5) and (3.6) in next sections.

3.3. Outliers in trend surface models

The usual way for estimating trend surface models (by iterated least squares), as well as the proposal by Crujeiras and Van Keilegom (2010) for non linear trends with spatial dependence, are not robust in the presence of outliers in the response. Now, taking into account the use of "pseudo-data", we propose a new procedure of trend surface models estimation which mitigates the effect of outliers in data.

Let's recall some notation. Consider a Gaussian spatial process $\{Z(\mathbf{s}), \mathbf{s} \in D\}$ and assume that the large-scale and small-scale behaviour of $Z(\mathbf{s})$ is described by the spatial regression model

$$Z(\mathbf{s}) = X(\mathbf{s})'\boldsymbol{\beta} + \varepsilon(\mathbf{s}), \quad \mathbf{s} \in D, \quad D \subset \mathbb{R}^d, \tag{3.4}$$

where $X(\cdot) \in \mathbb{R}^q$ denotes the spatial regressors given by geographic coordinates in trend surface models, and $m(\cdot) = X(\cdot)'\boldsymbol{\beta}$ is a linear trend component, which captures the large-scale variability of the process. $\varepsilon(\cdot)$ denotes the random component and shows the local behavior or small-scale evolution. The error process $\varepsilon(\cdot)$ is assumed to be zero-mean and second-order stationary, so in particular intrinsic stationary. Hence the dependence structure may be described from the variogram function, which is given by

$$2\gamma(\mathbf{u}) = \mathrm{Var}(\varepsilon(\mathbf{s}+\mathbf{u}) - \varepsilon(\mathbf{s})), \qquad \mathbf{u} \in \mathbb{R}^d, \;\; \mathbf{s}, \mathbf{s}+\mathbf{u} \in D.$$

Consider n locations $\{\mathbf{s}_1, \ldots, \mathbf{s}_n\}$ on the region D. The set of random variables corresponding with those locations will be represented by $\{Z(\mathbf{s}_1), \ldots, Z(\mathbf{s}_n)\}$ and the process will be described by the spatial regression model $Z(\mathbf{s}_i) = X(\mathbf{s}_i)\boldsymbol{\beta} + \varepsilon(\mathbf{s}_i)$, $i = 1, \ldots, n$. We denote $\mathbf{Z} = (Z(\mathbf{s}_1), \cdots, Z(\mathbf{s}_n))'$, $\mathbf{m} = (m(X(\mathbf{s}_1)), \ldots, m(X(\mathbf{s}_n)))'$.

Based on a sample, two sets of parameters must be estimated: $\boldsymbol{\beta} \in \mathbb{R}^q$, from the trend part and $\boldsymbol{\theta}$, the parameters from the dependence structure. Usually, $\boldsymbol{\theta}$ includes the range, the point variance (sill) and the nugget effect (microscale variation). Although the goal is the estimation of two sets of parameters, the estimation $\boldsymbol{\beta}$ is essential if we would like to do model based prediction.

The estimation procedure consists on four steps: (1) obtain a nonparametric estimator for the vector of observations $\hat{\mathbf{Z}} = (\hat{Z}(\mathbf{s}_1), \cdots, \hat{Z}(\mathbf{s}_n))$, considering $\hat{\mathbf{Z}} = \hat{\mathbf{m}}$, where $\hat{\mathbf{m}} = (\hat{m}(X(\mathbf{s}_1)), \ldots, \hat{m}(X(\mathbf{s}_n)))'$ (in the geostatistical context, $\hat{\mathbf{m}}$ may be written as $\hat{\mathbf{m}} = (\hat{m}(\mathbf{s}_1), \ldots, \hat{m}(\mathbf{s}_n))'$); (2) based on the "new response" $\hat{\mathbf{Z}}$ (pseudo-data), obtain an ordinary least squares estimator for $\boldsymbol{\beta}$, ignoring the dependence structure of the errors; (3) estimate the covariance matrix of the errors based on the residuals from the trend parameter estimator derived in step (2); (4) based on the "pseudo-data" $\hat{\mathbf{Z}}$, weighted least squares estimation of $\boldsymbol{\beta}$, taking the dependence structure of the errors, obtained in (3), into account.

We now explain each of these four steps in detail. First, a vector of pseudo-observations is constructed, $\hat{\mathbf{Z}} = (\hat{Z}(\mathbf{s}_1), \cdots, \hat{Z}(\mathbf{s}_n))' = (\hat{m}(\mathbf{s}_1), \ldots, \hat{m}(\mathbf{s}_n))'$. Here $\hat{m}(\mathbf{s}_i) = T(\hat{F}_i)$ (see Akritas (1996)), where $\hat{F}_i$ is an estimator of the conditional distribution function of $Z(\mathbf{s}_i)$ given the coordinates $\mathbf{s}_i$ and $T(\cdot)$ is a functional which for any distribution function F is determined by

$$T(F) = \int_0^1 F^{-1}(s)J(s)ds, \tag{3.5}$$

where $J(\cdot)$ is a given score function satisfying $\int_0^1 J(s)ds = 1$. Let $0 \leq \alpha_1 < \alpha_2 < \alpha_3 < \alpha_4 \leq 1$, with $\alpha_2 - \alpha_1 = \alpha_4 - \alpha_3$, and define the function $J(\cdot)$ as

$$J(s) = \frac{1}{\alpha_4 - \alpha_2} \begin{cases} 0 & s < \alpha_1 \\ 2\left(\dfrac{s-\alpha_1}{\alpha_2-\alpha_1}\right)^2 & \alpha_1 \leq s < \dfrac{\alpha_1+\alpha_2}{2} \\ 1 - 2\left(\dfrac{s-\alpha_2}{\alpha_1-\alpha_2}\right)^2 & \dfrac{\alpha_1+\alpha_2}{2} \leq s < \alpha_2 \\ 1 & \alpha_2 \leq s < \alpha_3 \\ 1 - 2\left(\dfrac{s-\alpha_3}{\alpha_4-\alpha_3}\right)^2 & \alpha_3 \leq s < \dfrac{\alpha_3+\alpha_4}{2} \\ 2\left(\dfrac{s-\alpha_4}{\alpha_4-\alpha_3}\right)^2 & \dfrac{\alpha_3+\alpha_4}{2} \leq s < \alpha_4 \\ 0 & \alpha_4 < s. \end{cases} \tag{3.6}$$

The score function is affected by the factor $1/(\alpha_4 - \alpha_2)$. Therefore, if the difference between α_4 and α_2 is small, then the function takes large values (or reciprocally). Moreover, it is symmetric about $(\alpha_1 + \alpha_4)/2$. Figure 3.1 plots this function considering different values of its parameters.

This function induces the value zero in the integrand of the expression (3.5) when $s < \alpha_1$ and $\alpha_4 < s$. If $\alpha_2 \leq s < \alpha_3$ then the contribution of the score function to the integrand would be constant.

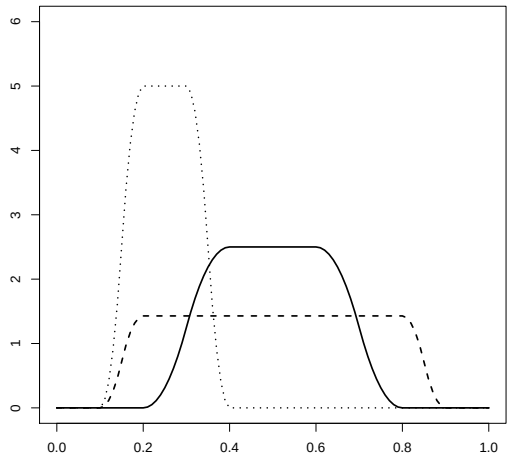

Figure 3.1: J score function with $\alpha_1 = 0.2$, $\alpha_2 = 0.4$, $\alpha_3 = 0.6$ and $\alpha_4 = 0.8$ (full line), with $\alpha_1 = 0.1$, $\alpha_2 = 0.2$, $\alpha_3 = 0.8$ and $\alpha_4 = 0.9$ (dashed line), and with $\alpha_1 = 0.1$, $\alpha_2 = 0.2$, $\alpha_3 = 0.3$ and $\alpha_4 = 0.4$ (dotted line).

Let $F_i(z) = F(z \mid \mathbf{s}_i)$ denote the conditional distribution function $Z(\mathbf{s}_i)$ given the coordinates $\mathbf{s}_i$.

Based on a sample of size n, let

$$\hat{F}_i(z) = \hat{F}(z \mid \mathbf{s}_i), \quad i = 1, \ldots, n,$$

denote the estimation of $F_i(\cdot)$. The estimation of the conditional distributional will be of the nearest neighbor kind. Following Stone (1977) consider estimators $\hat{F}_i(\cdot)$ of the form

$$\hat{F}_i(z) = \sum_{j=1}^{n} W_j(\mathbf{s}_i)\delta_{Z(\mathbf{s}_j)}, \quad i = 1, \ldots, n,$$

where $\delta_{Z(\mathbf{s}_j)} = \mathbb{I}_{Z(\mathbf{s}_j)\leq z}$ is a point mass at $Z(\mathbf{s}_j)$ and $W_j(\mathbf{s}_i)$ is a weight attached to the $j-$th observation out of the first n observations, which depends on $\mathbf{s}_1, \ldots, \mathbf{s}_n$ and on n but not on the attribute values. Let c_j, $1 \leq j \leq n$ be a triangular matrix of real numbers. The nearest neighbor weights are $W_j = c_{r(j)}$, where $r(j)$ is the rank of $\|\mathbf{s}_j - \mathbf{s}_i\|$. Nearest neighbors weights are called k nearest neighbor weights (k-nn) when for some k, $j > k$ implies $c_j = 0$. Another alternative for the estimation of the conditional distribution, in the current setting, is the consideration of a normal distribution, estimating the mean and standard deviation using a neighborhood criterion. This will be the option used in the simulations presented below.

Then the vector of pseudo-observations $\hat{\mathbf{Z}} = (\hat{Z}(\mathbf{s}_1), \cdots, \hat{Z}(\mathbf{s}_n))'$ will be $\hat{\mathbf{Z}} = \hat{\mathbf{m}}$, where $\hat{\mathbf{m}} = (\hat{m}(\mathbf{s}_1), \ldots, \hat{m}(\mathbf{s}_n))' = (T(\hat{F}_1), \ldots, T(\hat{F}_n))'$. This will be use as pseudo-response in next steps to mitigate the effect of outliers.

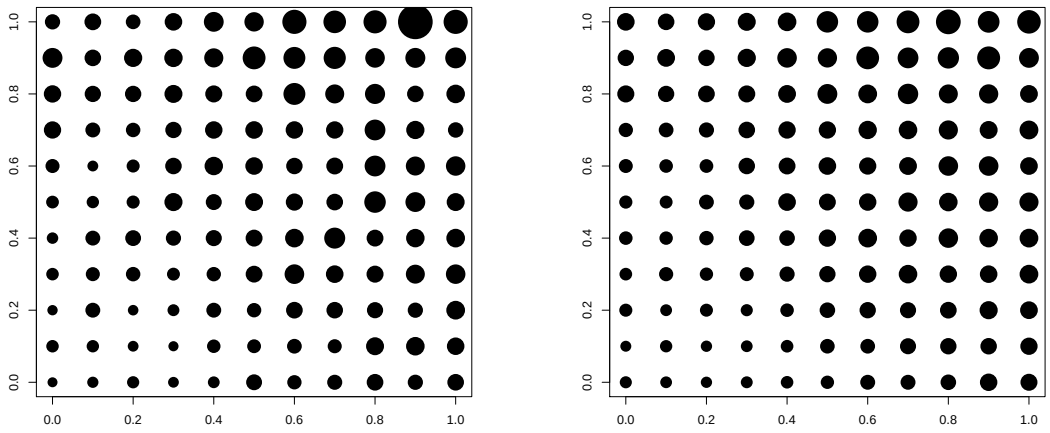

Figure 3.2: Sample data points with symbol area proportional to atributtes concentration (left) and to pseudo-data (right). Sample size: 121. Simulation results for a regular Gaussian spatial process whose trend is given by (3.2) with $\beta_0 = 2$, $\beta_1 = 1$ and $\beta_2 = 1$. The dependence structure is explained by a exponential covariogram (1.13) with parameters $c_e = 0.1$ and $a_e = 0.1$.

A sample of size 121 is simulated regularly from a Gaussian spatial process considering the unit square as support, whose trend is given by (3.2) with $\beta_0 = 2$, $\beta_1 = 1$ and $\beta_2 = 1$. The dependence structure is explained by a exponential covariogram (1.13) with parameters $c_e = 0.1$ and $a_e = 0.1$. The observation 120 is modified with the value of median+$3IQR$, where IQR denotes the interquartile range of the attributes. Figure 3.2 plots the sample points (left) and the corresponding pseudo-data (right). We should also notice that simulation of spatial processes with trend and dependence components may result in samples where both variability contributors may be confused. This fact complicates notably the design of simulation scenarios. In this case, if the range was chosen larger than one (hence it is

greater than any of the distances in the unit square), there would be a strong dependence. To avoid this, we consider range value 0.1. It can be seen that the new pseudo responses mitigate the effect of outliers existing in data. Note that $k = 4$ and $\alpha_1 = 0$, $\alpha_2 = 0.1$, $\alpha_3 = 0.9$ and $\alpha_4 = 1$ are chosen in the $J(\cdot)$ score function given in (3.1).

Now, using ordinary least squares using pseudo-data, from the linear regression of $\hat{\mathbf{Z}}$ over X, the vector $\boldsymbol{\beta}_{OLSP}$ is obtained as follows, ignoring the dependence structure of the errors. Here X denotes $X = (X(\mathbf{s}_1), \ldots, X(\mathbf{s}_n))'$.

$$\boldsymbol{\beta}_{OLSP} = \arg\min_{\boldsymbol{\beta}} (\hat{\mathbf{Z}} - X\boldsymbol{\beta})'(\hat{\mathbf{Z}} - X\boldsymbol{\beta}). \tag{3.7}$$

After obtaining the ordinary least squares estimation of $\boldsymbol{\beta}$, the estimation of the variance-covariance matrix of the errors is performed, based based on the residuals: $\hat{\epsilon}(\mathbf{s}_i) = Z(\mathbf{s}_i) - X(\mathbf{s}_i)\boldsymbol{\beta}_{OLSP}$, $i = 1, \ldots, n$. Given that $\varepsilon(\cdot)$ is a zero-mean and second-order stationary process, so in particular intrinsic stationary, the dependence structure may be described from the variogram function.

Given that errors are not used observed then the structure of dependence is estimated from the residuals of the regression $(\tilde{\varepsilon}(\mathbf{s}_1), \ldots, \tilde{\varepsilon}(\mathbf{s}_n))$. The classical non parametric estimator for the variogram is the empirical variogram (see Section 1.3), that for regression residuals is defined as

$$2\hat{\gamma}(\mathbf{u}) = \frac{1}{|N(\mathbf{u})|} \sum_{(i,j)\in N(\mathbf{u})} (\tilde{\varepsilon}(\mathbf{s}_i) - \tilde{\varepsilon}(\mathbf{s}_j))^2, \tag{3.8}$$

where $|N(\mathbf{u})|$ is the number of pairs in $N(\mathbf{u}) \equiv \{(i,j) : \mathbf{s}_i - \mathbf{s}_j = \mathbf{u}\}$. Cressie and Hawkins (1980) propose a more robust approach to the estimation of the variogram,

$$2\bar{\gamma}(\mathbf{u}) = \frac{1}{(0.457 + 0.494/|N(\mathbf{u})|)} \left\{ \frac{1}{|N(\mathbf{u})|} \sum_{(i,j)\in N(\mathbf{u})} |\tilde{\varepsilon}(\mathbf{s}_i) - \tilde{\varepsilon}(\mathbf{s}_j)|^{1/2} \right\}^4. \tag{3.9}$$

Variogram estimators are not generally valid, since they fail to satisfy the conditional negative definiteness property, or it is not easy to prove that this condition holds. However, nonparametric variogram estimators can be used as pilots for fitting a valid parametric model, by minimizing a certain criterion, such as least squares.

Suppose that a valid parametric variogram family is given by $\{2\gamma_{\boldsymbol{\theta}} : 2\gamma(\cdot) = 2\gamma(\cdot;\boldsymbol{\theta}), \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$. The parameter vector $\boldsymbol{\theta}$ can be estimated by using a weighted least squares approach, comparing the functions at lags $\mathbf{u}_1, \ldots, \mathbf{u}_K$ for some $K < \infty$:

$$\widehat{\boldsymbol{\theta}_{LS}} = \arg\min_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \hat{\Gamma}(\boldsymbol{\theta})' W(\boldsymbol{\theta}) \hat{\Gamma}(\boldsymbol{\theta}), \tag{3.10}$$

where $\hat{\Gamma}(\boldsymbol{\theta}) = (\gamma_{\boldsymbol{\theta}}(\mathbf{u}_1) - \hat{\gamma}(\mathbf{u}_1), \ldots, \gamma_{\boldsymbol{\theta}}(\mathbf{u}_K) - \hat{\gamma}(\mathbf{u}_K))'$ and $W(\boldsymbol{\theta})$ is an appropriate weight matrix. Ideally, $W(\boldsymbol{\theta}) = \mathrm{Cov}(\hat{\Gamma}(\boldsymbol{\theta}))^{-1}$ could be chosen, but it is not easy to find an expression for this covariance matrix, even for quite simple estimators of the variogram. When the errors would be observed and when the empirical estimator (3.8) or its robust version (3.9) is used, the expression for the covariance matrix of $\hat{\Gamma}(\boldsymbol{\theta})$ has been derived by Cressie (1985). It is a diagonal matrix with zero's everywhere except for variances of $2\hat{\gamma}(\mathbf{u}_l)$, $l = 1, \ldots, K$, on the diagonal, it is said $W = W(\boldsymbol{\theta}) = \mathrm{diag}\{\mathrm{var}[2\hat{\gamma}(\mathbf{u}_1)], \ldots, \mathrm{var}[2\hat{\gamma}(\mathbf{u}_1)]\}$. Note that, since we will use the response, without mitigating the outliers, a robust variogram estimator will be considered. A great simplification is obtained by assuming that dependence structures are functions only of the distance, that is to say, assume isotropy ($u = \|\mathbf{u}\|$). For choosing the number of lags K, the recommendations from Journel and Huijbregts (1978) are followed, with K fitting only up to half the maximum possible lag and considering only lags with $|N|(u)$ larger than 30, and $u_K \leq U/2$, where $U = \max\{\|\mathbf{u}\| : N(u) > 0\}$.

Now, since the process $\varepsilon(\mathbf{s})$ is second-order stationary, there is a function $C_{\boldsymbol{\theta}}(\cdot)$, the covariogram, such that:

$$C_{\boldsymbol{\theta}}(u) = \mathrm{Cov}(\varepsilon(\mathbf{s}), \varepsilon(\mathbf{s}+u)) = \sigma^2 - \gamma_{\boldsymbol{\theta}}(u) \tag{3.11}$$

which can be recovered from the variogram, and where σ^2 is the variance of the spatial process. If the second-order stationarity of the process does not hold, the covariogram $C_{\boldsymbol{\theta}}(\cdot)$ does not exists. In practice, this assumption can be checked using the test proposed by Fuentes (2005), whose approach is based on a spatial spectral analysis. Another problem is that sample covariances do not provide unbiased estimators of the underlying covariances. A valid estimator of the covariogram $C_{\boldsymbol{\theta}}(\cdot)$ can be obtained by plugging in (3.11) the corresponding estimator of the variogram and a suitable estimator $\hat{\sigma}^2$ of the variance. In most parametric variogram families, the variance parameter can be identified. In case this parameter can not be explicitly obtained from the model, then $(n-1)^{-1}\sum_{i=1}^{n}(\tilde{\varepsilon}(\mathbf{s}_i)-\bar{\tilde{\varepsilon}})^2$ may be used as an estimator of the variance, where $\bar{\tilde{\varepsilon}}$ denotes the average of the residuals.

Once the covariogram estimator has been obtained, consider the covariance matrix of the process $\{\varepsilon(\mathbf{s}_1),\ldots,\varepsilon(\mathbf{s}_n)\}$. This matrix is denoted by Σ and its entries are:

$$\Sigma(i,j) = C_{\boldsymbol{\theta}}(\mathbf{s}_i-\mathbf{s}_j), \qquad i,j=1\ldots,n.$$

This matrix can be estimated by $\hat{\Sigma}=(\hat{\Sigma}(i,j))_{i,j=1}^{n}$, where $\hat{\Sigma}(i,j)=C_{\widehat{\boldsymbol{\theta}_{LS}}}(\mathbf{s}_i-\mathbf{s}_j)$. Now, weighted least squares estimation of $\boldsymbol{\beta}$ is performed using pseudo-data, taking the dependence structure of the errors into account (contrary to the preliminary estimator $\boldsymbol{\beta}_{OLSP}$):

$$\boldsymbol{\beta}_{IGLSP} = \arg\min_{\boldsymbol{\beta}}(\hat{\mathbf{Z}}-X\boldsymbol{\beta})'\hat{\Sigma}^{-1}(\hat{\mathbf{Z}}-X\boldsymbol{\beta}). \tag{3.12}$$

Since $\hat{\Sigma}$ is an $n\times n$ symmetric and positive definite matrix, the Cholesky decomposition allows to write:

$$\hat{\Sigma}=\hat{L}\hat{L}',$$

where $\hat{L}$ is a lower triangular $n\times n$ matrix. Then the estimator of the regression parameter vector $\boldsymbol{\beta}_{IGLSP}$ can be written also as follows:

$$\boldsymbol{\beta}_{IGLSP} = \arg\min_{\boldsymbol{\beta}}\frac{1}{n}(\hat{L}^{-1}\hat{\boldsymbol{Z}}-\hat{L}^{-1}X\boldsymbol{\beta})'(\hat{L}^{-1}\hat{\boldsymbol{Z}}-\hat{L}^{-1}X\boldsymbol{\beta}).$$

Note that although in the proposed method we consider a least squares procedure for the estimation of the variogram, it could be replaced by any other pointwise consistent estimator.

A problem that arises is the construction of the third-step estimator for the dependence parameters, for instance, by a least squares criterion as in (3.10) based on $\hat{\boldsymbol{\varepsilon}}=\mathbf{Z}-X\boldsymbol{\beta}_{IGLSP}$. When estimating the variogram of a process based on residuals, the dependence structure in the data and the constraints required in the least squares procedure have to be taken into account. Therefore, the two-stage estimator of $\boldsymbol{\theta}$, obtained by replacing in the formula of $\widehat{\boldsymbol{\theta}_{LS}}$ the estimator $\boldsymbol{\beta}_{OLSP}$ by $\boldsymbol{\beta}_{IGLSP}$, will share with $\widehat{\boldsymbol{\theta}_{LS}}$ the same asymptotic properties, but the behavior on finite samples may not be satisfactory, see Gambolati and Galeati (1987).

3.4. Simulations

In order to study the performance of our procedure for trend surface models estimation with outliers, we will carry out a simulation study considering different scenarios. Firstly, we check the correct performance of the use of pseudo-data in simple and multiple linear regression, in presence of global outliers. This serves as a basis for proving the advantage of using pseudo-data to mitigate the effect of outliers. Finally, we will explore our proposed procedure with different configurations of large-scale (trend) and small-scale (dependence) variabilities.

It should be also noted that, along this section, two different procedures have been considered for obtaining pseudo-data. The proposal by Cristobal et al. (1987) has been applied in Section 3.4.1. with the aim of checking the effect without the presence of dependence. Nevertheless, the proposal by Akritas (1996) is considered 3.4.2. for trend surface models. Some discussion about this choice is presented in Section 3.5, jointly with some further issues which deserver a deeper investigation.

3.4.1. The use of pseudo-data in regression models

Consider a simple linear regression model:

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, \ldots, n, \tag{3.13}$$

being $\varepsilon_1, \ldots, \varepsilon_n \in N(0, \sigma^2)$, independent. In this case Y and X denote the response and the explanatory variables, respectively. Data from such a model will be generated, and an outlier will be introduced. A comparison between estimators obtained by ordinary least squares and by ordinary least squares using pseudo-data in the response is carried out. The pseudo-data are obtained from a Nadaraya-Watson regression estimator, as proposed by Cristobal et al. (1987). Bandwidth is selected by cross-validation.

A data sample of size 100 is simulated from a uniform distribution in $(0, 1)$, that would correspond with the explanatory variable X. The values of the theoretical coefficients will be $\beta_0 = 2$ and $\beta_1 = 1$, and the errors $\varepsilon_1, \ldots, \varepsilon_n \in N(0, 0.16)$. The response observation associated with the maximum value of the variable X (hence with a high leverage) is modified with the value median$(Y) + 3IQR$, where IQR denotes the interquartile range of the response variable, a value excessively high for the simulated sample, so this would be a global outlier and by the location of the X value also as an influential point.

Mean, median and mean squared error from 100 Monte Carlo experiments are reported for the trend estimators based on ordinary least squares (OLS) and ordinary least squares using pseudo-data (OLSP) in the Table 3.1. We use the notation OLS, OLSP and MSE from here on along the text. For both coefficients, it can be seen that estimates are better when OLSP is used and the MSE is relatively smaller using pseudo-data.

	OLS		OLSP	
	$\beta_0 = 2$	$\beta_1 = 1$	$\beta_0 = 2$	$\beta_1 = 1$
Mean	1.9847	1.0481	1.9952	1.0233
Median	1.9842	1.0538	1.9999	1.0247
MSE	0.0053	0.0185	0.0048	0.0154

Table 3.1: Sample size: 100. Simulation results for a simple linear regression (3.13) considering a uniform distribution in (0,1) as explanatory variable X, $\beta_0 = 2$ and $\beta_1 = 1$, and errors $N(0, 0.16)$. Mean, median and MSE from 100 Monte Carlo experiments are reported for the trend estimators based on ordinary least squares (OLS) and ordinary least squares using pseudo-data (OLSP).

Now, we check the correct performance of the use of pseudo data (in presence of outliers) from the coefficients regressors estimation in multiple linear regression models. The dimension $d = 2$ is considered, because it equates to the dimension of geographical coordinates in the spatial case. So, we consider a model

$$Y_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \varepsilon_i \quad i = 1, \ldots, n, \tag{3.14}$$

being $\varepsilon_1, \ldots, \varepsilon_n \in N(0, \sigma^2)$, independent. In this case Y is the response variable, while X_1 and X_2 are the explanatory variables. Again, we compare the estimators obtained by ordinary least squares and by ordinary least squares using pseudo-data in the response.

Two samples of size 100 are simulated from a uniform distribution in $(0, 1)$, which would correspond with the explanatory variables X_1 and X_2. The values of the theoretical coefficients will be $\beta_0 = 1$, $\beta_1 = 2$ and $\beta_2 = 1$, and the errors $\varepsilon_1, \ldots, \varepsilon_n \in N(0, 0.16)$. The response observation associated with the maximum of variables X_1 or X_2 is modified with the value median$(Y)+3IQR$, where IQR denotes the interquartile range of the response variable.

Mean, median and mean squared error from 100 Monte Carlo simulations are reported for the trend estimators based on ordinary least squares (OLS) and ordinary least squares using pseudo-data (OLSP) in the Table 3.2. For both coefficients, estimates are better when OLSP is used. For β_0, the MSE is relatively smaller if we use pseudo-data. However, for β_1 and β_2, the MSE is slightly smaller with usual procedure..

	OLS			OLSP		
	$\beta_0 = 1$	$\beta_1 = 2$	$\beta_2 = 1$	$\beta_0 = 1$	$\beta_1 = 2$	$\beta_2 = 1$
Mean	0.9699	2.0557	1.0613	0.9801	2.0520	1.0574
Median	0.9713	2.0738	1.0816	0.9804	2.0670	1.0803
MSE	0.0128	0.0386	0.0290	0.0125	0.0390	0.0302

Table 3.2: Sample size: 100. Simulation results for a multiple linear regression (3.14) considering uniform distributions in (0,1) as explanatory variables X_1 and X_2, $\beta_0 = 1$, $\beta_1 = 2$ and $\beta_2 = 1$, and errors $N(0, 0.16)$. Mean, median and MSE from 100 Monte Carlo experiments are reported for the trend estimators based on ordinary least squares (OLS) and ordinary least squares using pseudo-data (OLSP).

3.4.2. The use of pseudo-data in trend surface models

In order to explore the performance of our proposed method, we will carry out a simulation study considering different scenarios for the spatial regression model $Z(\mathbf{s}) = X(\mathbf{s})'\beta + \varepsilon(\mathbf{s})$, $\mathbf{s} \in D$, where $\mathbf{s} = (s_1, s_2)$. Remember that $X(\cdot) \in \mathbb{R}^q$ denotes the spatial regressors that are geographic coordinates in our case, and $m(\cdot) = X(\cdot)'\beta$ is a linear trend component, which captures the large-scale variability of the process. $\varepsilon(\cdot)$ denotes the random component and shows the local behavior or small-scale evolution.

Different comparisons in terms of the estimators of the linear trend component are made. Consider the estimator of $\boldsymbol{\beta}$ obtained using ordinary least squares (OLS) from the linear regression of $\mathbf{Z}$ over X. Analogous to those realized for the estimation given in (3.7), no type of dependence structure will be considered in the computation of this estimator. This vector, let us call it $\boldsymbol{\beta}_{OLS}$, is obtained as follows

$$\boldsymbol{\beta}_{OLS} = \arg\min_{\boldsymbol{\beta}} (\mathbf{Z} - X\boldsymbol{\beta})'(\mathbf{Z} - X\boldsymbol{\beta}). \tag{3.15}$$

The second estimator is the one which is usually computed in trend surface model estimation, following an iterative least squares procedure, which in the last step, yields a generalized least squares estimator based on estimated dependence parameters. This estimator is given by

$$\boldsymbol{\beta}_{IGLS} = \arg\min_{\boldsymbol{\beta}} (\mathbf{Z} - X\boldsymbol{\beta})'\hat{\Sigma}^{-1}(\mathbf{Z} - X\boldsymbol{\beta}). \tag{3.16}$$

Remember that the matrix $\hat{\Sigma}$ is an estimator of the covariance matrix of the process $\{\varepsilon(\mathbf{s}_1), \ldots, \varepsilon(\mathbf{s}_n)\}$, and its entries are $\hat{\Sigma}(i,j) = C_{\widehat{\boldsymbol{\theta}_{LS}}}(\mathbf{s}_i - \mathbf{s}_j)$, where $\widehat{\boldsymbol{\theta}_{LS}}$ are the dependence parameters estimated from OLS residuals.

We compare the estimator obtained based on ordinary least squares $\boldsymbol{\beta}_{OLS}$ (without use pseudo-data and without suppose any type of dependence structure), the estimator obtained based on ordinary least squares $\boldsymbol{\beta}_{OLSP}$ (using pseudo-data and without imposing any type of dependence structure), the estimator obtained based on generalized least squares $\boldsymbol{\beta}_{IGLS}$ (without using pseudo-data, but taking into account the dependence structure of the errors) and the estimator obtained based on generalized least squares $\boldsymbol{\beta}_{IGLSP}$ using pseudo-data and taking into account the dependence structure of the errors,

the one proposed in this work. Specifically, we compare the estimators given in (3.7), (3.12), (3.15) and (3.16).

Data samples with size 121 are generated from an isotropic Gaussian spatial process observed at regularly spaced locations $\{\mathbf{s}_1, \ldots, \mathbf{s}_n\}$ in the unit square. The first -order (planar) polynomial trend surface model for the mean of a two-dimensional process, where $\mathbf{s} = (s_1, s_2)$, is considered:

$$m(\mathbf{s}) = \beta_0 + \beta_1 s_1 + \beta_2 s_2. \tag{3.17}$$

Different values of the parameters β_0, β_1 and β_2 are chosen. We have also considered different isotropic dependence structures: spherical covariogram given in (1.11) and exponential covariogram given in (1.13), with different values of their parameters. No nugget effect is considered in the simulation scenarios. Moreover, for the compute the pseudo-data, $k = 4$ will be used to estimate the conditional distribution.

SCENARIO 1: $\boldsymbol{\beta} = (2, 1, 1)'$ and exponential covariogram with weak structure of dependence

Take $\beta_0 = 2$, $\beta_1 = 1$ and $\beta_2 = 1$, and assume that the dependence structure is modeled from a exponential covariogram (1.13) with parameters $c_e = 0.1$ and $a_e = 0.1$. The observation 120 (located in the top right corner of the square) is modified with the value of the own observation adding $3IQR$, where IQR denotes the interquartile range of the observed values of the process. In this case we choose the values of $\alpha_1 = 0$, $\alpha_2 = 0.1$, $\alpha_3 = 0.9$ and $\alpha_4 = 1$ in the score function $J(\cdot)$ given in (3.1).

	OLS			OLSP		
	$\beta_0 = 2$	$\beta_1 = 1$	$\beta_2 = 1$	$\beta_0 = 2$	$\beta_1 = 1$	$\beta_2 = 1$
Mean	1.9374	1.0716	1.0865	1.9970	1.0057	1.0113
Median	1.9405	1.0807	1.0765	1.9963	0.9965	0.9978
MSE	0.0278	0.0439	0.0374	0.0245	0.0398	0.0302

	IGLS			IGLSP		
	$\beta_0 = 2$	$\beta_1 = 1$	$\beta_2 = 1$	$\beta_0 = 2$	$\beta_1 = 1$	$\beta_2 = 1$
Mean	1.9439	1.0635	1.0884	2.0032	1.0013	1.0065
Median	1.9548	1.0881	1.0895	1.9963	0.9915	0.9984
MSE	0.0241	0.0367	0.0365	0.0245	0.0365	0.0278

Table 3.3: Sample size: 121. Simulation results for a regular Gaussian spatial process whose trend is given by (3.18) with $\beta_0 = 2$, $\beta_1 = 1$ and $\beta_2 = 1$. The dependence structure is explained by a exponential covariogram (1.13) with parameters $c_e = 0.1$ and $a_e = 0.1$. Mean, median from 100 Monte Carlo experiments are reported for the trend estimators based on ordinary least squares (OLS) and ordinary least squares using pseudo-data (OLSP), generalized least squares without using pseudo-data (IGLS) and generalized least squares using pseudo-data (IGLSP).

Table 3.3 presents mean, median and MSE from 100 Monte Carlo experiments for the trend estimators based on ordinary least squares (OLS), ordinary least squares using pseudo-data (OLSP), generalized least squares(IGLS) and generalized least squares using pseudo-data (IGLSP). All estimates are better when we use OLSP instead of OLS, and the MSE is smaller if we use pseudo-data. If we consider the dependence structure, estimates are better when we use IGLSP instead of IGLS. Finally, a comparison between OLSP and IGLSP is realized. Again, all estimates are better when we use IGLSP instead of OLSP, and the MSE is smaller if we consider the dependence structure too. As a conclusion, in all comparisons estimates obtained from IGLSP are better. Note that, although we do not show the estimations of parameters of the variogram in the Table 3.3, they have been also taken into account. If we do not consider the pseudo-data, the mean of 100 Monte Carlo outcomes of the variance is 0.0988, while it is 0.0983 if we consider pseudo-data. As for the estimation of the range, its estimate is 0.2150, being smaller 0.2131 if we use pseudo-data. Recall that the theoretical values are $c_e = 0.1$ and $a_e = 0.1$, so in both cases the estimation of the range is not good.

Histograms of the IGLS and IGLSP estimates are plotted in the Figure 3.3, in top and bottom row, respectively. Just by looking and the plots, it is clear that estimates (IGLS and IGLSP) seemed centred around the theoretical values.

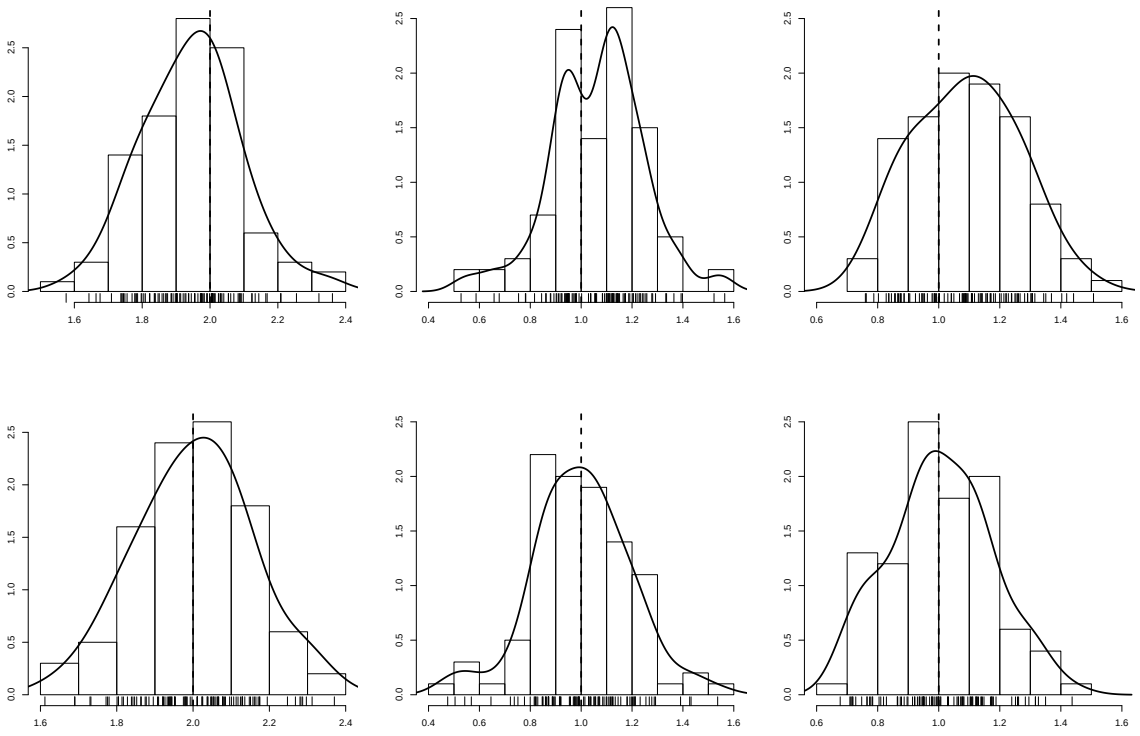

Figure 3.3: Histograms of the estimates of IGLS (toprow) and IGLSP (bottonrow) with Gaussian kernel density estimator (the method of Sheather and Jones to select the bandwidth is used), from 100 Monte Carlo simulations of a Gaussian spatial process (sample size: 121) with exponential covariogram (1.13) and parameters $c_e = 0.1$ and $a_e = 0.1$; and $\beta_0 = 2$, $\beta_1 = 1$ and $\beta_2 = 1$.

Now, a sample of size 400 is simulated regularly from a Gaussian spatial process considering the same trend and the same covariance function than in the previous simulations. The observation 400 (located in the top right corner of the square) is modified with the value of the observation plus $3IQR$, where IQR denotes the interquartile range of the observed values of the process. Again, we choose the

values of $\alpha_1 = 0$, $\alpha_2 = 0.1$, $\alpha_3 = 0.9$ and $\alpha_4 = 1$ in the score function $J(\cdot)$ given in (3.1).

Results for this scenario are reported in Table 3.4. It can be observed that, in general, the mean squared error of all estimators reduces with respect to the results with sample size 121. The IGLSP estimator provides better results also in this case . Regarding the dependence parameter, the mean of 100 Monte Carlo outcomes of the variance is 0.0925, no matter if pseudo-data are considered. For the estimation of the range, its estimate is 0.1967 with raw data, being smaller 0.1964 if we use pseudo-data. For the range, it should be noted that better results are obtained when increasing the sample size.

	OLS			OLSP		
	$\beta_0 = 2$	$\beta_1 = 1$	$\beta_2 = 1$	$\beta_0 = 2$	$\beta_1 = 1$	$\beta_2 = 1$
Mean	1.9674	1.0399	1.0382	1.9934	1.0091	1.0085
Median	1.9767	1.0370	1.0294	2.0007	1.0210	0.9933
MSE	0.0225	0.0358	0.0355	0.0217	0.0352	0.0339

	IGLS			IGLSP		
	$\beta_0 = 2$	$\beta_1 = 1$	$\beta_2 = 1$	$\beta_0 = 2$	$\beta_1 = 1$	$\beta_2 = 1$
Mean	1.8932	1.1372	1.1303	2.0093	0.9959	0.9916
Median	1.8978	1.1202	1.1214	2.0007	1.0002	0.9855
MSE	0.0332	0.0491	0.0539	0.0217	0.0280	0.0297

Table 3.4: Sample size: 400. Simulation results for a regular Gaussian spatial process which trend is given by (3.18) with $\beta_0 = 2$, $\beta_1 = 1$ and $\beta_2 = 1$. The dependence structure is explained by a exponential covariogram (1.13) with parameters $c_e = 0.1$ and $a_e = 0.1$. Mean, median from 100 Monte Carlo experiments are reported for the trend estimators based on ordinary least squares (OLS) and ordinary least squares using pseudo-data (OLSP), generalized least squares (IGLS) and generalized least squares using pseudo-data (IGLSP).

SCENARIO 2: $\boldsymbol{\beta} = (2, 1, 1)'$ and spherical covariogram with weak dependence structure

This new scenario is a modified version of the previous one: instead of the exponential covariogram, the spherical covariogram given in (1.11) is considered (with the same value of parameters). Note that the same parameters for the trend component are chosen too. Again, we choose the values of $\alpha_1 = 0$, $\alpha_2 = 0.1$, $\alpha_3 = 0.9$ and $\alpha_4 = 1$ in the score function $J(\cdot)$ given in (3.1).

Mean, median and MSE from 100 Monte Carlo experiments are reported for the four estimators, as shown the Table 3.5. In this case, the use of pseudo-data mitigate the effect of outliers for all coefficients of trend component (we are comparing OLS with OLSP and IGLS with IGLSP). Furthermore, if we consider the dependence structure, we may observe that the estimates obtained by IGLS and IGLSP are similar than those obtained by OLS and OLSP, respectively. The same is true for the MSE. In all the comparisons, estimates obtained from IGLSP are better.

	OLS			OLSP		
	$\beta_0 = 2$	$\beta_1 = 1$	$\beta_2 = 1$	$\beta_0 = 2$	$\beta_1 = 1$	$\beta_2 = 1$
Mean	1.9360	1.0751	1.0884	1.9951	1.0111	1.0129
Median	1.9401	1.0775	1.0881	2.0015	1.0038	1.0074
MSE	0.0097	0.0152	0.0149	0.0059	0.0101	0.0074

	IGLS			IGLSP		
	$\beta_0 = 2$	$\beta_1 = 1$	$\beta_2 = 1$	$\beta_0 = 2$	$\beta_1 = 1$	$\beta_2 = 1$
Mean	1.9360	1.0753	1.0893	1.9956	1.0108	1.0128
Median	1.9402	1.0798	1.0863	2.0015	1.0039	1.0075
MSE	0.0098	0.0152	0.0152	0.0059	0.0101	0.0074

Table 3.5: Sample size: 121. Simulation results for a regular Gaussian spatial process which trend is given by (3.18) with $\beta_0 = 2$, $\beta_1 = 1$ and $\beta_2 = 1$. The dependence structure is explained by a spherical covariogram (1.11) with parameters $c_s = 0.1$ and $a_s = 0.1$. Mean, median from 100 Monte Carlo experiments are reported for the trend estimators based on OLS, and ordinary least squares using pseudo-data (OLSP), generalized least squares (IGLS) and generalized least squares using pseudo-data(IGLSP).

SCENARIO 3: $\boldsymbol{\beta} = (2, 1, 1)'$ and exponential covariogram with higher variance

Now, consider that the dependence structure is modeled again by an exponential covariogram (1.13), but the value of the variance is three times larger, $c_e = 0.3$. The values of the parameters of the trend component chosen are the same that in the above cases, $\beta_0 = 2$, $\beta_1 = 1$ and $\beta_2 = 1$. The observation 120 (top right corner of the square) is modified with the value of the observation plus $3IQR$, where IQR denotes the interquartile range of the observed values of the process. In this case we choose the values of $\alpha_1 = 0.3$, $\alpha_2 = 0.4$, $\alpha_3 = 0.5$ and $\alpha_4 = 0.6$ in the score function $J(\cdot)$ given in (3.1).

Table 3.6 shows mean, median and MSE from 100 Monte Carlo outcomes. The correct performance of the use of pseudo-data may be observed again (if we compare OLS with OLSP and IGLS with IGLSP). Taking into account the dependence structure of the errors, IGLS provides best results than OLS. As for IGLSP, this method leads to better estimates than OLSP. In all the comparisons, estimates obtained from IGLSP are better. If we do not consider the pseudo-data, the mean of 100 Monte Carlo replies of the variance is 0.2820, while it is 0.2806 if we consider them. As for the estimation of the range, its estimate is 0.1156, being smaller (0.1150) if we use pseudo-data.

	OLS			OLSP		
	$\beta_0 = 2$	$\beta_1 = 1$	$\beta_2 = 1$	$\beta_0 = 2$	$\beta_1 = 1$	$\beta_2 = 1$
Mean	1.9190	1.0960	1.1103	1.9258	1.0169	1.0139
Median	1.9271	1.1022	1.1118	1.9371	1.0118	0.9967
MSE	0.0230	0.0370	0.0328	0.0225	0.0294	0.0218

	IGLS			IGLSP		
	$\beta_0 = 2$	$\beta_1 = 1$	$\beta_2 = 1$	$\beta_0 = 2$	$\beta_1 = 1$	$\beta_2 = 1$
Mean	1.9192	1.0963	1.1113	1.9261	1.0168	1.0139
Median	1.9303	1.1042	1.1128	1.9371	1.0126	0.9967
MSE	0.0230	0.0370	0.0334	0.0225	0.0293	0.0219

Table 3.6: Sample size: 121. Simulation results for a regular Gaussian spatial process which trend is given by (3.18) with $\beta_0 = 2$, $\beta_1 = 1$ and $\beta_2 = 1$. The dependence structure is explained by a exponential covariogram (1.13) with parameters $c_e = 0.3$ and $a_e = 0.1$. Mean, median from 100 Monte Carlo experiments are reported for the trend estimators based on ordinary least squares (OLS) and ordinary least squares using pseudo-data (OLSP), generalized least squares (IGLS) and generalized least squares using pseudo-data (IGLSP).

Now, a sample of size 400 is simulated regularly from a Gaussian spatial process considering the same trend and the same covariance function than in the previous simulation. The observation 400 (located in the top right corner of the square) is modified with the value of the observation plus $3IQR$, where IQR denotes the interquartile range of the observed values of the process. Again, we choose the values of $\alpha_1 = 0.3$, $\alpha_2 = 0.4$, $\alpha_3 = 0.5$ and $\alpha_4 = 0.6$ in the score function $J(\cdot)$ given in (3.1).

Table 3.7 shows median and MSE from 100 Monte Carlo experiments. Once again, it can be seen that the MSE of all estimators is smaller when the sample size is larger. The same conclusion can be drawn regarding the comparison of the estimates of β_1 and β_2. However, we cannot conclude the same for the estimation of β_0. In this case, the best results seems to be obtained by OLS, which indeed requires further investigation.

The mean of 100 Monte Carlo replies of the variance is 0.3000, not affecting to the use or not of pseudo-data. As for the estimation of the range, its estimate is 0.1030, the same for both cases too. Better conclusion can be drawn regarding the behaviour for increasing sample size.

We have consider that the dependence structure is stronger in terms of the variance. However, we may also consider that the dependence structure is stronger from the range. If we consider that the dependence structure is explained by a exponential covariogram (1.13) with parameters $c_e = 0.1$ and $a_e = 0.3$ instead of by a exponential covariogram (1.13) with parameters $c_e = 0.3$ and $a_e = 0.1$, then the results are similar, for this reason we omit them. Note that, in this case we should chose the parameters $\alpha_1 = 0$, $\alpha_2 = 0.05$, $\alpha_3 = 0.95$ and $\alpha_4 = 1$ in the $J(\cdot)$ function.

	OLS			OLSP		
	$\beta_0 = 2$	$\beta_1 = 1$	$\beta_2 = 1$	$\beta_0 = 2$	$\beta_1 = 1$	$\beta_2 = 1$
Mean	1.9629	1.0447	1.0474	1.9328	1.0077	1.0098
Median	1.9639	1.0464	1.0434	1.9276	1.0227	1.0138
MSE	0.0127	0.0196	0.0209	0.0166	0.0189	0.0192

	IGLS			IGLSP		
	$\beta_0 = 2$	$\beta_1 = 1$	$\beta_2 = 1$	$\beta_0 = 2$	$\beta_1 = 1$	$\beta_2 = 1$
Mean	1.9416	1.0734	1.0717	1.9371	1.0050	1.0047
Median	1.9399	1.0669	1.0756	1.9276	1.0121	1.0118
MSE	0.0145	0.0216	0.0245	0.0166	0.0183	0.0187

Table 3.7: Sample size: 400. Simulation results for a regular Gaussian spatial process which trend is given by (3.18) with $\beta_0 = 2$, $\beta_1 = 1$ and $\beta_2 = 1$. The dependence structure is explained by a exponential covariogram (1.13) with parameters $c_e = 0.3$ and $a_e = 0.1$. Mean, median from 100 Monte Carlo experiments are reported for the trend estimators based on ordinary least squares (OLS) and ordinary least squares using pseudo-data (OLSP), generalized least squares (IGLS) and generalized least squares using pseudo-data (IGLSP).

SCENARIO 4: $\boldsymbol{\beta} = (2, 3, 2)'$ and exponential covariogram with a weak dependence structure

Finally, consider that the dependence structure is modeled by the exponential covariogram (1.13) with parameters $c_e = 0.1$ and $a_e = 0.1$. The values of the parameters of the trend component chosen are larger that in the above cases, $\beta_0 = 2$, $\beta_1 = 3$ and $\beta_2 = 2$. The observation 120 (top right corner of the square) is modified with the value of the observation plus $3IQR$, where IQR denotes the interquartile range of the observed values of the process. In this case we choose the values of $\alpha_1 = 0.2$, $\alpha_2 = 0.3$, $\alpha_3 = 0.5$ and $\alpha_4 = 0.6$ in the score function $J(\cdot)$ given in (3.1).

Results are shown in Table 3.8. The correct performance of the pseudo-data may be observed again (if we compare OLS with OLSP and IGLS with IGLSP). Taking into account the dependence structure of the errors, for all estimates IGLSP provides best results than IGLS. As for IGLSP, this method leads to better estimates than OLSP for β_0, but not for β_1 and β_2, although results are quite similar. This is because, the dependence structure is weaker and the values of the parameters of the trend component are large, therefore the dependence structure is harder to capture. Note that, these differences are small. Moreover, if we do not consider the pseudo-data, the mean of 100 Monte Carlo outcomes of the variance is 0.1123, while it is smaller if we consider them, 0.1100. As for the estimation of the range, its estimate is 0.1216, being also smaller if we use pseudo-data, 0.1189.

	OLS			OLSP		
	$\beta_0 = 2$	$\beta_1 = 3$	$\beta_2 = 2$	$\beta_0 = 2$	$\beta_1 = 3$	$\beta_2 = 2$
Mean	1.8523	3.1708	2.2079	1.9528	2.9179	1.9740
Median	1.8586	3.1723	2.1994	1.9584	2.9167	1.9627
MSE	0.0275	0.0391	0.0501	0.0079	0.0166	0.0080

	IGLS			IGLSP		
	$\beta_0 = 2$	$\beta_1 = 3$	$\beta_2 = 2$	$\beta_0 = 2$	$\beta_1 = 3$	$\beta_2 = 2$
Mean	1.8518	3.1713	2.2108	1.9551	2.9149	1.9727
Median	1.8577	3.1755	2.2001	1.9584	2.9154	1.9624
MSE	0.0277	0.0392	0.0515	0.0079	0.0170	0.0081

Table 3.8: Sample size: 121. Simulation results for a regular Gaussian spatial process which trend is given by (3.18) with $\beta_0 = 2$, $\beta_1 = 3$ and $\beta_2 = 2$. The dependence structure is explained by a exponential covariogram (1.13) with parameters $c_e = 0.1$ and $a_e = 0.1$. Mean, median from 100 Monte Carlo experiments are reported for the trend estimators based on ordinary least squares (OLS) and ordinary least squares using pseudo-data (OLSP), generalized least squares (IGLS) and generalized least squares using pseudo-data (IGLSP).

Now, a sample of size 400 is simulated regularly from a Gaussian spatial process considering the same trend and the same covariance function than in the previous simulation. The observation 400 (located in the top right corner of the square) is modified with the value of the observation plus $3IQR$, where IQR denotes the interquartile range of the observed values of the process. Again, we choose the values of $\alpha_1 = 0.3$, $\alpha_2 = 0.4$, $\alpha_3 = 0.5$ and $\alpha_4 = 0.6$ in the score function $J(\cdot)$ given in (3.1).

Table 3.9 shows mean, median and MSE from 100 Monte Carlo experiments. Once again, it can be seen that the MSE of all estimators is smaller when the sample size is bigger. The same conclusion can be drawn regarding the comparison of the estimates of β_0, β_1 and β_2 when increasing the sample size.

If we do not consider the pseudo-data, the mean of 100 Monte Carlo outcomes of the variance is 0.1017, while it is smaller if we consider them, giving 0.1015. As for the estimation of the range, its estimate is 0.1032, which continues being also smaller than if we use pseudo-data 0.1030.

	OLS			OLSP		
	$\beta_0 = 2$	$\beta_1 = 3$	$\beta_2 = 2$	$\beta_0 = 2$	$\beta_1 = 3$	$\beta_2 = 2$
Mean	1.9411	3.0717	2.0733	1.9337	2.9781	1.9917
Median	1.9387	3.0728	2.0704	1.9299	2.9864	1.9908
MSE	0.0073	0.0111	0.0116	0.0084	0.0068	0.0065

	IGLS			IGLSP		
	$\beta_0 = 2$	$\beta_1 = 3$	$\beta_2 = 2$	$\beta_0 = 2$	$\beta_1 = 3$	$\beta_2 = 2$
Mean	1.9032	3.1202	2.1192	1.9440	2.9660	1.9828
Median	1.9036	3.1116	2.1159	1.9299	2.9717	1.9830
MSE	0.0133	0.0202	0.0209	0.0084	0.0072	0.0065

Table 3.9: Sample size: 400. Simulation results for a regular Gaussian spatial process which trend is given by (3.18) with $\beta_0 = 2$, $\beta_1 = 1$ and $\beta_2 = 1$. The dependence structure is explained by a exponential covariogram (1.13) with parameters $c_e = 0.1$ and $a_e = 0.1$. Mean, median from 100 Monte Carlo experiments are reported for the trend estimators based on ordinary least squares (OLS) and ordinary least squares using pseudo-data (OLSP), generalized least squares (IGLS) and generalized least squares using pseudo-data (IGLSP).

3.5. Some discussion and open problems

To perform our proposed of trend surface model estimation with outliers, it is necessary to go through different steps detailed in the Section 3.3. Each one of these steps contains some estimations. In this section we review possible not fixed parameters that may have an impact in the estimation procedure, such as those used to compute the pseudo-data. We also contemplate some open problems including: some extensions of the procedure and the use of our proposal in kriging, among others.

Tuning parameters for computing the pseudo-data

To obtain the pseudo-data (remember that they are obtained in a non parametric way), it is necessary to estimate the expression given in (3.5), which implies using a score function $J(\cdot)$ and the estimation of the conditional distribution function.

- **The selection of alpha for the J function**

  In the simulation study performed in the Section 3.4.2., for each particular simulation, we have considered fixed values for the parameters α_1, α_2, α_3 and α_4. The choice of these parameters may influence in the procurement of pseudo-data. Let us see an example.

  Considering a sample size 121 regularly simulated from a Gaussian spatial process whose trend is given by (3.18) with $\beta_0 = 2$, $\beta_1 = 1$ and $\beta_2 = 1$, with exponential covariogram (1.13) and parameters $c_e = 0.1$ and $a_e = 0.1$. As usual, the observation 120 is modified with the value adding

$3IQR$. Figure 3.4 shows scatterplots of the pseudo-data against data, considering different values for the parameters of the score function $J(\cdot)$: $\alpha_1 = 0$, $\alpha_2 = 0.1$, $\alpha_3 = 0.9$ and $\alpha_4 = 1$ (left); $\alpha_1 = 0.3$, $\alpha_2 = 0.4$, $\alpha_3 = 0.6$ and $\alpha_4 = 0.7$ (center); and $\alpha_1 = 0.3$, $\alpha_2 = 0.35$, $\alpha_3 = 0.4$ and $\alpha_4 = 0.7$ (right). The pseudo-data depend on each value of α_1, α_2, α_3 or α_4, but they directly depend on the center of symmetry $(\alpha_1 + \alpha_4)/2$, see Figure 3.4 (left) and (center). The pseudo-data obtained seem to be the same (they are different in the fourth decimal), even though the value of the parameters are different. Bearing this issue in mind, the choice of the parameters α_2 and α_3 is important, because they provide a translation of the data, see Figure 3.4 (right).

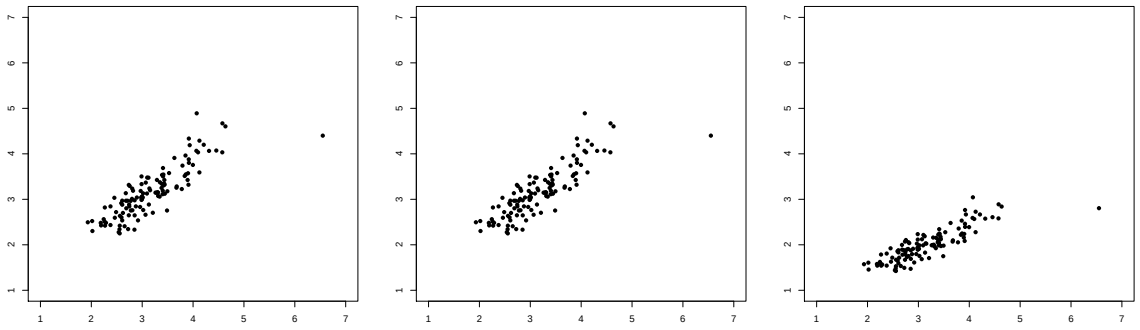

Figure 3.4: Sample size: 121. Regularly simulated data from a Gaussian spatial process whose trend is given by (3.18) with $\beta_0 = 2$, $\beta_1 = 1$ and $\beta_2 = 1$, with exponential covariogram (1.13) and parameters $c_e = 0.1$ and $a_e = 0.1$. Scatterplots of the pseudo-data versus data considering different values for the parameters of the score function $J(\cdot)$. $\alpha_1 = 0$, $\alpha_2 = 0.1$, $\alpha_3 = 0.9$ and $\alpha_4 = 1$ (left), $\alpha_1 = 0.3$, $\alpha_2 = 0.4$, $\alpha_3 = 0.6$ and $\alpha_4 = 0.7$ (center) and $\alpha_1 = 0.3$, $\alpha_2 = 0.35$, $\alpha_3 = 0.4$ and $\alpha_4 = 0.7$ (right).

	IGLSP		
$\boldsymbol{\alpha}'$	$\beta_0 = 2$	$\beta_1 = 1$	$\beta_2 = 1$
$(0, 0.1, 0.9, 1)'$	2.0032	1.0013	1.0064
$(0.3, 0.4, 0.6, 0.7)'$	2.0032	1.0013	1.0064
$(0.3, 0.35, 0.4, 0.7)'$	1.2766	0.6287	0.6300

Table 3.10: Sample size: 121. Simulation results for a regular Gaussian spatial process which trend is given by (3.18) with $\beta_0 = 2$, $\beta_1 = 1$ and $\beta_2 = 1$. The dependence structure is explained by a exponential covariogram (1.13) with parameters $c_e = 0.1$ and $a_e = 0.1$. Average values from 100 Monte Carlo experiments are reported for the trend estimator based on generalized least squares using pseudo-data (IGLSP) considering different values for the parameters of the score function $J(\cdot)$: $\alpha_1 = 0$, $\alpha_2 = 0.1$, $\alpha_3 = 0.9$ and $\alpha_4 = 1$ (left); $\alpha_1 = 0.3$, $\alpha_2 = 0.4$, $\alpha_3 = 0.6$ and $\alpha_4 = 0.7$ (center); and $\alpha_1 = 0.3$, $\alpha_2 = 0.35$, $\alpha_3 = 0.4$ and $\alpha_4 = 0.7$ (right). The vector $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \alpha_3, \alpha_4)'$.

Table 3.10 shows mean from 100 Monte Carlo experiments from the IGLSP estimator, considering different values for the parameters of the score function $J(\cdot)$. It can be seen that the values of

the estimates in the two first cases are the same. As mentioned above, differences are found from the fourth decimal figure. We omit the MSE of each estimate because they are similar. For the last case, the estimates are clearly more biased (downwards for all them), and its MSE is larger. This issue requires further investigation.

- **The selection of k for the conditional distribution**

  The number of k-nearest neighbors chosen to perform the estimation of the conditional distribution function may also influence the pseudo-data. If we consider many k-nearest neighbors then we are not very accurate with the estimation. That is because in spatial data close observations tend to be more similar than observations which are far apart. Therefore, if we consider many k-nearest neighbors then we may chose observations which are not very similar or whose dependence is not relevant. We should not chose too few either.

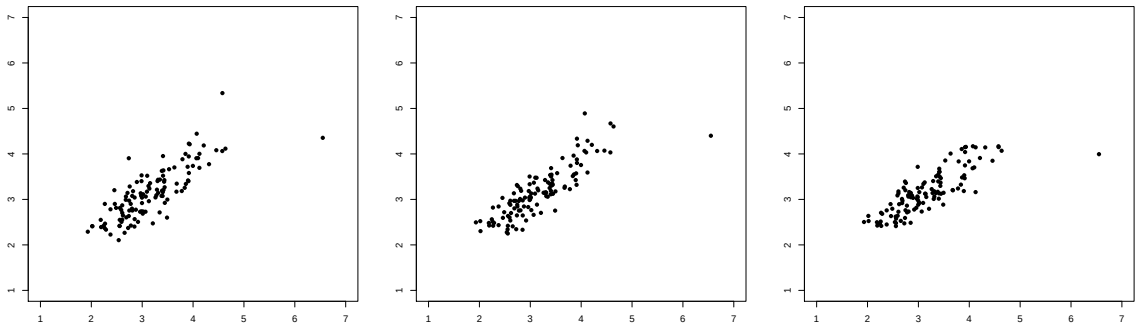

Figure 3.5: Sample size: 121. Regularly simulated data from a Gaussian spatial process whose trend is given by (3.18) with $\beta_0 = 2$, $\beta_1 = 1$ and $\beta_2 = 1$, with exponential covariogram (1.13) and parameters $c_e = 0.1$ and $a_e = 0.1$. Scatterplots of the data versus pseudo-dataconsidering different k-nearest neighbors: $k = 2$ (left), $k = 4$ (center) and $k = 20$ (right).

	IGLSP		
k	$\beta_0 = 2$	$\beta_1 = 1$	$\beta_2 = 1$
2	1.9796	1.0106	1.0017
4	2.0032	1.0013	1.0065
20	2.1720	0.8306	0.8367

Table 3.11: Sample size: 121. Simulation results for a regular Gaussian spatial process which trend is given by (3.18) with $\beta_0 = 2$, $\beta_1 = 1$ and $\beta_2 = 1$. The dependence structure is explained by a exponential covariogram (1.13) with parameters $c_e = 0.1$ and $a_e = 0.1$. Average values from 100 Monte Carlo experiments are reported for the trend estimator based on generalized least squares using pseudo-data (IGLSP) considering different values of k.

In the simulation study, we have considered $k = 4$. Given that simulations were realized regularly in the unit squared, we are chosen as nearest neighbors the north, south, east and west of each

observation, except which are in the boundary of the unit squared. If we considerer for example $k = 2$, then the estimates obtained by our procedure are not better. The same applies when, for example $k = 20$ is chosen. Below, the effect of the k will be shown graphically.

With a sample size of 121, considering the unit square as support, data have been drawn regularly taking the exponential model given in (1.13) as covariance function with $c_e = 1$ and $a_e = 0.1$, and whose trend is given by (3.18) with $\beta_0 = 2$, $\beta_1 = 1$ and $\beta_2 = 1$. The observation 120 is modified with the value of the observation plus $3IQR$. Figure 3.5 shows scatterplots of these data versus pseudo-data, considering different k-nearest neighbors chosen to realize the estimation of the conditional distributional function, $k = 2$ (left), $k = 4$ (center) and $k = 20$ (right).

The estimates of IGLSP considering different values of k are shown in the Table 3.11. It can be observed that the estimates are clearly more biased when $k = 20$ is considered (upwards for β_0 and downwards for β_1 and β_2).

Some improvements on the variogram estimation

In the method proposed. the variogram is computed based on residuals from a regression model. As it is pointed out in Kim and Boos (2004), the empirical variogram based on residuals is seriously biased downwards, implying an underestimation of the variance when considering the method in (3.10). Figure 3.6 shows histograms of the variance estimation (using pseudo-data) with Gaussian kernel density estimator (the method of Sheather and Jones to select the bandwidth is used). 100 Monte Carlo simulations from a Gaussian spatial process (sample size: 121 (left) and 400 (right)) with exponential covariogram (1.13) and parameters $c_e = 0.1$ and $a_e = 0.1$; and $\beta_0 = 2$, $\beta_1 = 1$ and $\beta_2 = 1$. The authors monotonize the empirical variogram, applying the pool adjacent violators algorithm (see Kim and Boos (2004)).

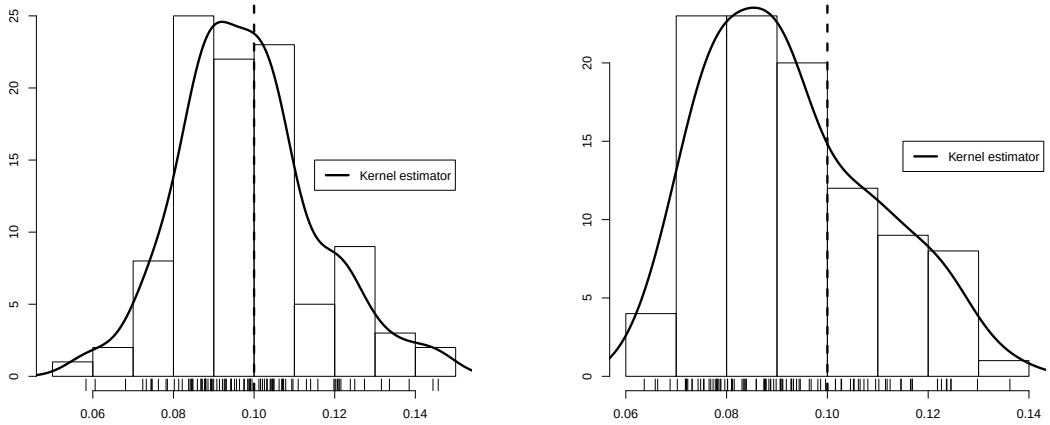

Figure 3.6: Histograms of the estimates of the variance estimation (using pseudo-data) from 100 Monte Carlo simulations of a Gaussian spatial process (sample size: 121 (left) and 400 (right)) with exponential covariogram (1.13) and parameters $c_e = 0.1$ and $a_e = 0.1$; and $\beta_0 = 2$, $\beta_1 = 1$ and $\beta_2 = 1$.

The case of irregularly spaced data

In the Section 3.4.2, all simulations were performed regularly from a Gaussian spatial process considering as support the unit square. However, data may be drawn irregularly. We have also explore some scenarios under irregular design, and one of them is reported in Table 3.12. At first sight, it can be noticed that results are considerably worse than for the simulation scenarios presented in the

previous section. Results from OLS and IGLS are quite bad, whereas OLSP and IGLSP see, to try to capture only the global effect β_0.

This issue may seem to pose a drawback for the practical application of the proposal. However, it should be mentioned that data coming from satellite measures are regularly spaced, and there are many examples of such data. Nevertheless, the adaptation of the method for irregularly sampled data is required. We could also like to point out a that a cornerstone here is the dependence parameters. If we do not consider the pseudo-data, the mean of 100 Monte Carlo outcomes of the variance is 0.2836, while it is 0.2878 if we consider pseudo-data. As for the estimation of the range, its estimate is 0.0432, being larger 0.0436 if we use pseudo-data. Recall that the theoretical values are $c_e = 0.1$ and $a_e = 0.1$, so in both cases the estimation of the variance and the range are not good.

	OLS			OLSP		
	$\beta_0 = 2$	$\beta_1 = 1$	$\beta_2 = 1$	$\beta_0 = 2$	$\beta_1 = 1$	$\beta_2 = 1$
Mean	3.0067	-0.0081	0.0335	1.9959	0.0183	0.0272
Median	2.9835	-0.0381	0.0393	2.0038	0.0458	0.0372
MSE	1.0582	1.0846	1.0184	0.0953	1.1334	1.0990

	IGLS			IGLSP		
	$\beta_0 = 2$	$\beta_1 = 1$	$\beta_2 = 1$	$\beta_0 = 2$	$\beta_1 = 1$	$\beta_2 = 1$
Mean	3.0082	-0.0093	0.0332	1.9913	0.0247	0.0252
Median	2.9887	-0.0247	0.0318	2.0038	0.0437	0.0337
MSE	1.0619	1.0899	1.0202	0.0953	1.1142	1.1043

Table 3.12: Sample size: 121. Simulation results for a irregular Gaussian spatial process which trend is given by (3.18) with $\beta_0 = 2$, $\beta_1 = 1$ and $\beta_2 = 1$. The dependence structure is explained by a exponential covariogram (1.13) with parameters $c_e = 0.1$ and $a_e = 0.1$. Mean, median from 100 Monte Carlo experiments are reported for the trend estimators based on ordinary least squares (OLS) and ordinary least squares using pseudo-data (OLSP), generalized least squares (IGLS) and generalized least squares using pseudo-data (IGLSP).

The use of a local linear smoother

In the simulation study carried out for trend surface models, we have considered the proposal given by Akritas (1996) to compute the pseudo-data. Although a kernel-type smoother could be used (imitating Cristobal et al. (1987)). The proposal by Akritas (1996) is more general since it can be adapted to the incomplete data case (for handling censoring and truncation). Nevertheless, to illustrate the use of a kernel estimator, we repeat all the scenarios considering the local linear estimator. Similar results, but more biased (downwards for β_0 and upwards for β_1 and β_2), have been observed. We considered the cross-validation bandwidth for smoothing. Note that, we have also tried with other bandwidth values, obtaining similar results.

Some extensions of the procedure

Our procedure is intended for linear trend surface with outliers, however, some extensions could be obtained.

A first modification is the one that arises for non linear trends. So the idea would be to modified the proposal by Crujeiras and Van Keilegom (2010) which is not robust in the presence of outliers in the response.

A second modification would be to consider a explanatory variable in the linear trend (and not only the geographic coordinates). Obviously, non linear trends and spatial covariates in the model could be introduced.

The use of our proposal in kriging

Kriging is an interpolation method for spatial prediction, which is not robust to the presence of outliers. Hence, the use of pseudo-data directly on kriging interpolation could be beneficial. In addition, our proposal could be quite useful for the particular case of universal kriging or for kriging with external drift.

Some notes about the theoretical properties

Consider a Gaussian spatial process $\{Z(\mathbf{s}), \mathbf{s} \in D\}$ and assume that the large-scale and small-scale behaviour of $Z(\mathbf{s})$ is described by the spatial regression model given in (3.4). Consider n locations $\{\mathbf{s}_1, \ldots, \mathbf{s}_n\}$ on the region D, the set of random variables corresponding with those locations will be represented by $\{Z(\mathbf{s}_1), \ldots, Z(\mathbf{s}_n)\}$. Therefore, the process will be described by the spatial regression model $Z(\mathbf{s}_i) = X(\mathbf{s}_i)\boldsymbol{\beta} + \varepsilon(\mathbf{s}_i)$, $i = 1, \ldots, n$. Note that, the errors $(\varepsilon(\mathbf{s}_1), \ldots, \varepsilon(\mathbf{s}_n)) \sim N(0, \Sigma)$, where $\Sigma(i,j) = C_{\boldsymbol{\theta}}(\mathbf{s}_i - \mathbf{s}_j), i, j = 1 \ldots, n$.

In order to prove the asymptotic properties of the estimator $\boldsymbol{\beta}_{IGLSP}$, given that we are working in a spatial setting, we must first describe how this asymptotic framework evolves. This is needed to guarantee the consistency of the dependence parameters estimators as proved by Lahiri et al. (2002).

The asymptotic framework required is known as "shrinking asymptotics", a combination of increasing domain and infilling asymptotics. It is needed for consistent estimators of $\boldsymbol{\theta}$ parameter.

Denote by R_0 an open subset of $(-1/2, 1/2]^2$ containing the origin (so that the shape is preserved by inflation) and consider λ_n a sequence of real numbers such that $\lambda_n \to \infty$ and $n \to \infty$. The prototype sampling region is given by:

$$R_n = \lambda_n R_0.$$

Consider now $\Delta = \text{diag}(\delta_1, \delta_2)$ and $\mathcal{Z}^2 = \{\Delta\mathbf{i}, \mathbf{i} \in \mathbb{Z}^2\} = \{(\delta_1 i_1, \delta_2 i_2), \mathbf{i} \in \mathbb{Z}^2\}$ (the integer lattice in $\mathbb{R}^2$ has an increment δ_1 in the horizontal direction and an increment δ_2 in the vertical direction. For the sake of simplicity, we will take $\delta_1 = \delta_2 = 1$. If the observations of the random process are taken at $\{\mathbf{s}, \mathbf{s} \in \mathcal{Z}^2 \cap R_n\}$ this is a pure increasing domain asymptotic framework.

For the mixed asymptotic framework (shrinking asymptotics), consider also a a sequence $h_n \to 0$ as $n \to \infty$. The sampling points are given by $\{\mathbf{s}1, \ldots, \mathbf{s}_n\} = \{\mathbf{s}, \mathbf{s} \in h_n\mathcal{Z}^2 \cap R_n\}$. In this case, the scaled lattice $h_n\mathcal{Z}^2$ becomes finer as h_n tends to zero (as n increases).

Always under the previous framework , and in order to aplly Lahiri et al. (2002) results for the estimators of the dependence parameters, $\widehat{\boldsymbol{\theta}_{LS}}$, the following condition must be satisfy by $\boldsymbol{\beta}_{OLSP}$ (obtained with pseudo-data):

$$\lambda_n^2 \|\boldsymbol{\beta}_{OLSP} - \boldsymbol{\beta}\|^4 = o_{\mathbb{P}}(1).$$

This condition implies that $\boldsymbol{\beta}_{OLSP}$ converges to $\boldsymbol{\beta}$ at a rate faster than $\lambda_n^{-1/2}$. This result will hold if we prove the asymptotic normality with rate $\sqrt{n}$. Regarding the properties of $\boldsymbol{\beta}_{OLSP}$, in the usual regression setting, Akritas (1996) proved that

$$\sqrt{n}(\boldsymbol{\beta}_{OLS} - \boldsymbol{\beta}) \to N(\mathbf{0}, \Sigma),$$

so this result should be extended to our asymptotic framework.

With the previous arguments, we would have $\widehat{\boldsymbol{\theta}_{LS}}$, which is a strong consistent estimator of $\boldsymbol{\theta}$. Finally, for obtaining asymptotic properties of $\boldsymbol{\beta}_{IGLSP}$, we may write this estimator as

$$\boldsymbol{\beta}_{IGLSP} = \arg\min_{\boldsymbol{\beta}} \|M(\boldsymbol{\beta}, \hat{\mathbf{Z}}, \widehat{\boldsymbol{\theta}_{LS}})\|,$$

under certain conditions for M. Properties of such an estimator (but just with a single nuisance) have been studied by Chen et al. (2003). In our case, we have to two nuisances: a non-parametric nuisance in the pseudo-data and a parametric one in the estimated covariance matrix. So the results of Chen et al. (2003) should be extended.

Nevertheless, this is only an indication to perform the theoretical properties, we have not prove them yet.

The development of proper statistical theory for the estimators will enable the quantification of their uncertainty. In fact, this is something that is missing in the real application presented in the next section. The lack of theoretical results hampers the assessment of significance for the trend estimators. In addition, it should be noted that in the spatial setting, resampling techniques for this type of problems, have not been successfully introduced in the statistical iterature. This is mainly due to the fact that, in the resampling scheme, one should imitate both large scale and small scale variability components.

3.6. Illustration with real data

In order to illustrate the performance of our method, we consider in this section a real data set. This data set, obtained from Gomez and Hazen (1970), collects coal ash for the Robena Mine Property in Green County, Pennsylvania, which is available in the `R` package `gstat`. These data come from the Pittsburgh coal seam that is associated with a deltaic sedimentation system that includes much of southwestern Pennsylvania, northwestern Ohio and northern West Virginia. The 208 coal-ash core measurements at locations with west coordinates greater than 64000 feet are considered. This defines an approximately square grid, with 2500 feet spacing, running southwest to northeast and northwest to southeast. Figure 3.7 (left) plots sample data points for coal ash with symbols area proportional to measured concentration. It seems that the observation 60 which is in the location $(5, 6)$ could be an spatial outlier. Note that the variogram cloud and the Moran scatterplot detect this observation as outlier.

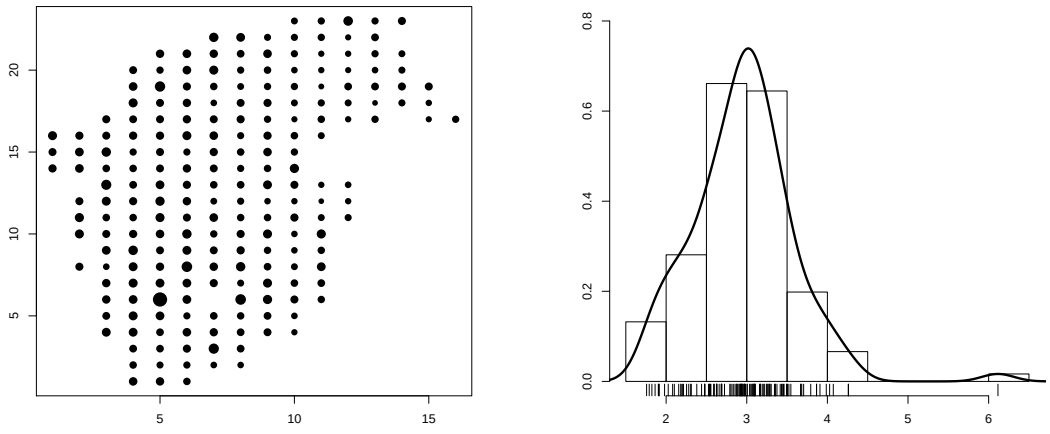

Figure 3.7: Sample data points for coal ash with symbols area proportional to measured concentration (left) and histogram of the measures concentration (right).

An histogram of this concentration with Gaussian kernel density estimator (the method of Sheather and Jones to select the bandwidth is used) is plotted in Figure 3.7 (right). In Figure 3.8, we show the scatterplots of the coal ash against the coordinates, in the East-West and North-South directions. This representation reveals the presence of linear trend in the East-West direction, but little or not trend in the North-South direction. From this analysis, we will check the performance of our procedure considering the first-order (planar) polynomial trend surface model for the mean of a two-dimensional process, $\mathbf{s} = (s_1, s_2)$:

$$m(\mathbf{s}) = \beta_0 + \beta_1 s_1 + \beta_2 s_2. \tag{3.18}$$

where s_1 and s_2 are, respectively, the coordinates in the East-West and North-South directions.

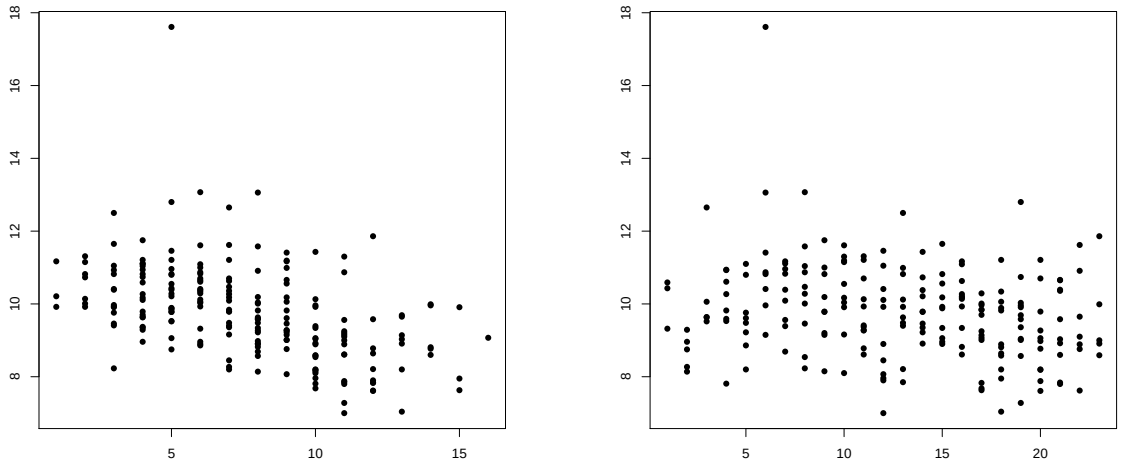

Figure 3.8: Coal ash against cordinates. Left plot: East-West direction. Right plot: North-South direction.

We consider different values of the paramaters of the $J(\cdot)$ function, and we denote with a superscript each of them (described in Table 3.13 caption). Table 3.13 shows the estimates for the trend based on ordinary least squares (OLS), ordinary least squares using pseudo-data (OLSP), generalized least squares(IGLS) and different generalized least squares using pseudo-data. The estimates obtained based on OLS and OLSP do not take into account the structure of dependence of the errors. Moreover, according with Cressie (1993) we know that the coal ash data exhibit a strong linear trend in the East-West direction but there is no trend in the North-South direction. Therefore the estimates β_2 should be approximately zero.

We may see that the results obtained by different procedures are quite similar. Unfortunately, we cannot asses the significance of the estimates for β_2, which would be interesting for this particular example.

The results indicate that, as we move from West to Esat, the concentration of coal ash is reduced, but again, we should quantify the uncertainty of the estimates obtained.

Note that, if we do not consider the pseudo-data, the variance estimated is 0.9971, while it is 0.9969 if we consider pseudo-data. As for the estimation of the range, its estimate is 1.2991, being smaller 1.2988 if we use pseudo-data.

	β_0	β_1	β_2
OLS	11.2468	-0.1771	-0.0104
OLSP	11.2544	-0.1669	-0.0160
IGLS	11.2121	-0.1756	-0.0085
IGLSP^1	11.2193	-0.1648	-0.0147
IGLSP^2	11.8865	-0.1543	-0.0298
IGLSP^3	10.5914	-0.1777	-0.0011
IGLSP^4	10.0167	-0.1892	0.0113

Table 3.13: estimates for the trend based on ordinary least squares (OLS), ordinary least squares using pseudo-data (OLSP), generalized least squares(IGLS) and different generalized least squares using pseudo-data: IGLSP^1 if $\alpha_1 = 0$, $\alpha_2 = 0.1$, $\alpha_3 = 0.9$ and $\alpha_4 = 1$; by IGLSP^2 if $\alpha_1 = 0.4$, $\alpha_2 = 0.6$, $\alpha_3 = 0.8$ and $\alpha_4 = 1$; by IGLSP^3 if $\alpha_1 = 0$, $\alpha_2 = 0.2$, $\alpha_3 = 0.4$ and $\alpha_4 = 0$.; and by IGLSP^4 if $\alpha_1 = 0$, $\alpha_2 = 0.1$, $\alpha_3 = 0.2$ and $\alpha_4 = 0.3$

References

Akritas, M. G. (1996). On the use of nonparametric regression techniques for fitting parametric regression models. *Biometrics*, **52**, 1342-1362.

Anselin, L. (1995). Local indicators of spatial association-LISA. *Geographical analysis*, **27**(2), 93-115.

Anselin, L. (1996). The Moran scatterplot as an ESDA tool to assess local instability in spatial association. *Spatial analytical perspectives on GIS*, **111**, 111-125.

Anselin, L. (2005). *Exploring spatial data with GeoDaTM: a workbook.* Urbana-Champaign, IL: University of Illinois, Center for Spatially Integrated Social Science, Spatial Analysis Laboratory.

Barnett, V., and Lewis, T. (1994). *Outliers in Statistical Data.* John Wiley & Sons, New York.

Beckman, R.J., and Cook, R.D. (1983). Outliers. *Technometrics*, **25**(2), 119-149.

Bowman, A. W., and Crujeiras, R. M. (2013). Inference for variograms. *Computational Statistics & Data Analysis*, **66**, 19-31.

Chen, X., Linton, O., and Van Keilegom, I. (2003). Estimation of semiparametric models when the criterion function is not smooth. *Econometrica*, **71**(5), 1591-1608.

Chiles, J. P., and Delfiner, P. (1999). *Geostatistics: Modeling Spatial Uncertainty.* John Wiley & Sons, New York.

Cliff, A. D., and Ord, J. K. (1973). *Spatial autocorrelation.* Pion, London.

Cliff, A. D., and Ord, J. K. (1981). *Spatial processes: models & applications.* Pion, London.

Cressie, N. (1985). Fitting variogram models by weighted least squares. *Journal of the International Association for Mathematical Geology*, **17**(5), 563-586.

Cressie, N. (1993). *Statistics for spatial data.* John Wiley & Sons, New York.

Cressie, N., and Hawkins, D.M. (1980). Robust estimation of the variogram: I. *Journal of the International Association for Mathematical Geology*, **12**(2), 115-125.

Cristobal, J. C., Roca, P. F., and Manteiga, W. G. (1987). A class of linear regression parameter estimators constructed by nonparametric estimation. *The Annals of Statistics*, **15**(2), 603-609.

Crujeiras, R.M., and Van Keilegom, I. (2010). Least squares estimation of nonlinear spatial trends. *Computational Statistics & Data Analysis*, **54**(2), 452-465.

Diggle, P., and Ribeiro, P.J. (2007). *Model-based Geostatistics.* Springer, New York.

Fernández, J.A., Real, C., Couto, J.A., Aboal, J.R. and Carballeira, A. (2005). The effect of sampling design on extensive bryomonitoring surveys of air pollution. *Science of the total environment*, **337**(1), 11-21.

Fuentes, M. (2005). A formal test for nonstationarity of spatial stochastic processes. *Journal of Multivariate Analysis*, **96**(1), 30-54.

Gallant, A. R., and Goebel, J. J. (1976). Nonlinear regression with autocorrelated errors. *Journal of the American Statistical Association*, **71**(356), 961-967.

Gambolati, G., and Galeati, G. (1987). Comment on "Analysis of Nonintrinsic Spatial Variability by Residual Kriging with Application to Regional Groundwater Level" by Shlomo P. Neuman and Elizabeth A. Jacobson. *Mathematical Geology*, **19**(3), 249-257.

Geary, R. C. (1954). The contiguity ratio and statistical mapping. *The Incorporated Statistician*, **5**(3), 115-146.

Getis, A., and Ord, J. K. (1992). The analysis of spatial association by use of distance statistics. *Geographical Analysis*, **24**(3), 189-206.

Gomez, M., and Hazen, K. (1970). *Evaluating sulfur and ash distribution in coal seams by statistical response surface regression analysis (No. BM-RI-7377).* Bureau of Mines, Denver, Colo.(USA).

Guan, Y., Sherman, M., and Calvin, J. A. (2004). A nonparametric test for spatial isotropy using subsampling. *Journal of the American Statistical Association*, **99**(467), 810-821.

Hawkins, D.M. (1980). *Identification of Outliers.* Chapman and Hall, London.

Journel, A.G., and Huijbregts, C.J. (1978). *Mining Geostatistics.* Academic press, London.

Jowett, G. H. (1952). The accuracy of systematic sampling from conveyor belts. *Applied Statistics*, **1**, 50-59.

Kim, H.J., and Boos, D.D. (2004). Variance estimation in spatial regression using nonparametric semivariogram based on residuals. *Scandinavian Journal of Statistics*, **31**, 387-401.

Lahiri, S. N., Lee, Y., and Cressie, N. (2002). On asymptotic distribution and asymptotic efficiency of least squares estimators of spatial variogram parameters. *Journal of Statistical Planning and Inference*, **103**(1), 65-85.

Lindgren, B. (1976). *Statistical theory.* Macmillan, New York.

Lu, N., and Zimmerman, D. L. (2005). Testing for directional symmetry in spatial dependence using the periodogram. *Journal of Statistical Planning and Inference*, **129**(1), 369-385.

Mandelbrot, B. B., and Van Ness, J. W. (1968). Fractional Brownian motions, fractional noises and applications. *SIAM review*, **10**(4), 422-437.

Mantel, N. (1967). The detection of disease clustering and a generalized regression approach. *Cancer Research*, **27**(2), 209-220.

Matheron, G. (1962). *Traité de Géostatistique Appliquée, Tome I. Mémoires du bureau de recherches géologiques et minières.* Editions Technip, Paris.

Matheron, G. (1971). *The Theory of Regionalized Variables and its Applications.* Les Cahiers de Morphologie Mathematique, Fontainebleau.

Moran, P. A. (1948). The interpretation of statistical maps. *Journal of the Royal Statistical Society. Series B (Methodological)*, **10**(2), 243-251.

Mörters, P., and Peres, Y. (2010). *Brownian Motion.* Cambridge University Press.

Ord, J. K., and Getis, A. (1994). Distributional issues concerning distance statistics. Working paper.

Shekhar, S., Lu, C.T. and Zhang, P. (2003). A unified approach to detecting spatial outliers. *GeoInformatica*, **7**(2), 139-166.

Stone, C. J. (1977). Consistent nonparametric regression. *The Annals of Statistics*, 595-620.

Yaglom, A. M. (1957). Some classes of random fields in n-dimensional space, related to stationary random processes. *Theory of Probability & Its Applications*, **2**(3), 273-320.

Zimmerman, D. L., and Stein, M. (2010). Classical geostatistical methods. *Handbook of Spatial Statistics*, 29-44.